Ontsluitingsdocument Pyttersen's-project

Onderzoeksteam:

Aansturing en initiatie: Joost Berkhout (d.j.berkhout@uva.nl), Marcel Hanegraaff en Eelke Heemskerk

Assistentie: Roderick Witjas en Max Boiten

Dit is een Klein Data Project (KDP) gefinancierd door de DANS-KNAW (2018).

We bedanken voormalig redacteur Alice Garritsen voor haar bereidheid om het redactionele werk toe te lichten. Verder gaat onze dank uit naar kopieerrechtenhouder Thorald Wevers voor zijn welwillendheid om ruimte te geven aan het academische gebruik van de historische almanakgegevens.

UNIVERSITEIT VAN AMSTERDAM

AMSTERDAM INSTITUTE FOR SOCIAL SCIENCE RESEARCH

Inhoud

Inleiding

Dit document omschrijft de achtergrond, het proces en de gegevensstructuur van de databestanden van het 'Pyttersen's almanak' project. Dit project is uitgevoerd aan de Universiteit van Amsterdam bij de afdeling politicologie gedurende 2018. Er zijn verschillende vroege edities van de Pyttersen's almanak gedigitaliseerd en in een databestand samengebracht. De jaarlijkse bestanden van de recentere CD-ROM gegevens zijn samengevoegd.

De Pyttersen's almanak is een naslagwerk van de organisatorische opmaak van het maatschappelijk middenveld. Er staan ook gegevens in over bestuurlijke organen en andere organisaties in de (semi)-publieke sector. De almanak is regelmatig gebruikt voor politicologisch onderzoek, bijvoorbeeld in het klassieke werk van Lijphart over Verzuiling (1968), het 'Graven naar Macht' onderzoek, de AIAS Dataset Brancheorganisaties (http://www.uva-aias.net/nl/branche-organisatie), en in recent internationaal vergelijkend survey onderzoek (cigsurvey.eu). De informatie uit deze reeks almanakken is waardevol voor verschillende wetenschappelijke disciplines en is de beschikbaarheid in de vorm van een gegevensbestand is van maatschappelijk belang.

Voorafgaand aan dit project, werd het gebruik van gegevens uit de almanakken serie bemoeilijkt doordat de informatie uit vroege edities alleen op papier beschikbaar is en de informatie van de cd-roms van latere edities niet tussen de jaren gekoppeld is. Daarnaast is het gebruik van deze informatie relatief omslachtig omdat de almanak niet meer wordt uitgegeven en daardoor in afnemende mate beschikbaar is via bibliotheken. Dit vergroot de kans dat deze waardevolle informatie verloren gaat. Met het oog op deze ontwikkelingen zijn in dit project de beschikbare digitale gegevens (1997-2008) op een systematische manier georganiseerd in een longitudinaal bestand, en zijn vijf oude papieren edities (1956, 1971, 1985 en 1992-1993) gescand en in een set van gegevens bestanden omgezet. Bij de Koninklijke Bibliotheek bestaat het voornemen om de edities tussen 1901 en 1958 te digitaliseren. Dit project geldt dan als een pilot voor het omzetten van scans in gegevensbestanden. Dit kan in een vervolgproject relatief eenvoudig voor alle edities gedaan worden.

Het werk samenvattend, zijn er vijf papieren edities van de almanak gedigitaliseerd en samengevoegd tot een dataset. Verder is ook de data van de cd-roms samengevoegd in een dataset. Dit zijn ongeveer 15000 entries per editie. Deze datasets zijn samen met de scans van de almanakken beschikbaar gemaakt. In dit document wordt dit proces kort besproken, evenals de specificaties van de data, de keuzes die zijn gemaakt gedurende dit proces en hoe de opgeleverde data gebruikt kan worden.

Bronbeschrijving

De Pyttersen's almanak is de enige bron in Nederland die een historisch en compleet overzicht geeft van het maatschappelijk middenveld en de politiek-ambtelijke organisatie. De almanak bevat informatie over particuliere, overheids- en semi-overheidsorganen en –instellingen die op zijn minst een supralokaal belang dienen en primair een non-profit doel nastreven in de Nederlandse samenleving. De almanak was decennia lang de adressen-'Bijbel' van personen en organisaties met enige politiek, bestuurlijke of maatschappelijke relevantie en werd intensief gebruikt door politiek journalisten, lobbyisten, ambtenaren en politici[1].

De eerste uitgave van de almanak was in 1901, samengesteld door Hendrik Tjeerds Pyttersen, lid van de Tweede Kamer der Staten Generaal. De laatste editie van de Pyttersen's almanak is verschenen in 2013. Nadien heeft de uitgever, Bohn Stafleu van Loghum, de rechten overgedragen aan Thorald Wevers (thoroldwevers@hotmail.com). Vanaf jaargang 1997 werd de data ook op een cd-rom bijgeleverd. Door de jaren heen is het werk veranderd van een echte almanak, met daarin tabellen over bijvoorbeeld oude lengtematen of telegraafkosten, naar een naslagwerk voor verenigingen. Zo is in de loop van de jaren '50 is de focus meer komen te liggen op de almanak als 'adressen-Bijbel' en verdween ook de reclame uit de almanak. De indeling van de almanak verandert door de jaren heen op twee manieren: enerzijds verandert de inhoud van de almanak in reactie op ontwikkelingen in maatschappelijke of organisatorische structuur (zoals het onafhankelijk worden van Suriname); anderzijds verandert de indeling van de almanak zelf. Dit betekent dat de almanakken een overzicht geven van de maatschappelijke en bestuurlijke veranderingen in de Nederlandse samenleving en kunnen dus als waardevolle bron voor onderzoek dienen. Het uiteenvallen van de verzuiling is bijvoorbeeld zichtbaar in de verandering van de samenstelling van belangenverenigingen: van een indeling gebaseerd op vier zuilen naar het samengaan van deze organisaties in een koepelvereniging op landelijk niveau.

De almanak werd vormgegeven naar de al bestaande 'Staatsalmanak', maar dan bedoeld voor een breder publiek en met aandacht voor allerlei soorten verenigingen in de Nederlandse maatschappij die een boven lokale interesse. De inhoud van de almanak kwam tot stand doordat, ten eerste, organisaties uit zichzelf zich meldde ter opname in de almanak, maar, meestal, ten tweede, doordat de redactie organisaties benaderde met het verzoek om gegevens te verstrekken voor de almanak. Organisaties werden benaderd wanneer ze in het nieuws kwamen, in relevante registers opgenomen waren (bv bij ministeries) of op een andere manier in het openbare leven deelnemen. Ieder jaar

[1] zie bv: https://www.volkskrant.nl/politiek/nationaal-erfgoed-de-pyttersen~a4219055/
en: A.J. van Beelen en T. N. M. Schuyt (2001) *Een Nationaal Wereldboek. 100 jaar Pyttersen's Nederlandse Almanak.* Houten: Bohn Stafleu Van Loghum.

werden organisaties verzocht mutaties in de samenstelling of de contactgegevens te melden. In de (latere) almanakken staan zo 15.000 organisaties die op onderwerp zijn gerubriceerd, waar bij iedere organisatie namen van contactpersonen, adressen en eventuele publicaties worden genoemd. Met de tijd is deze informatie bijgewerkt en is, naast een telefoon- of faxnummer, in latere edities ook een e-mailadres of webpagina toegevoegd. Ook is er een uitgebreid alfabetische register toegevoegd om het zoeken te vergemakkelijken. De enorme hoeveelheid aan data die in deze almanakken staat, is interessant voor politicologisch of sociologisch onderzoek. Deze data biedt de mogelijk voor het onderzoeken van bijvoorbeeld de veranderingen in het maatschappelijke middenveld over tijd, maar ook voor onderzoek over netwerken en bestuurssamenstellingen.

Data

Dit project omvat de volgende bestanden:

- '97-08.json' en 97-08.CSV
- 'boeken.json' en boeken.CSV
- 5 almanak scans in PDF van de edities: 1956, 1971, 1985 en 1992-1993
- TXT bestanden met de categorie-structuur per editie
- Dit ontsluitingdocument en codeboek met een omschrijving van de variabelen in de csv-bestanden

De twee hoofddatasets zijn '97-08.json' en 'boeken.json' in het bestandsformaat '*.json*' (JavaScript Object Notation). '97-08.json' bevat de samengevoegde data van de cd-roms voor de jaren 1997-2009 en 'boeken.json' bevat de combineerde data van vier gescande en verwerkte almanakken voor de jaren 1956, 1971, 1985 en 1992-1993. In deze bestanden staat de data uit de almanakken zoveel mogelijk in originele vorm en hieruit kan data voor onderzoek worden onttrokken.

Van de almanakken zijn er vanaf 1997 ook losse CSV-bestanden per jaar gemaakt. Deze bestanden zijn beknopter en bedoeld om de data te kunnen gebruiken zonder dat eerst data uit de .json bestanden hoeft worden te onttrokken. Zie het codeboek voor een overzicht van de variabelen in deze bestanden. Tot slot zijn er vijf pdf's van de gescande almanakken per jaar. Deze worden bijgeleverd, zodat de mogelijkheid bestaat om de dataset te checken.

Proces

Het doel van dit project was om een dataset te creëren waarin data uit verschillende papieren almanakken en data van cd-roms staan. Dit proces omvat verschillende stappen die hieronder nader zullen worden beschreven. Eerst zal de bronnenselectieprocedure worden beschreven. De volgende

stap, het digitaliseren van de almanakken, zal ook kort worden beschreven. Tot slot wordt uiteengezet hoe de twee datasets (97-08.json en boeken.json) uit de gedigitaliseerde data zijn geconstrueerd en welke keuzes hier aan ten grondslag liggen.

Selectieprocedure

De volgende edities zijn gescand en gedigitaliseerd: 1956, 1971, 1985, 1992-1993 en 1998-1999.[2] Er is dus voor ongeveer ieder decennium een almanak geselecteerd. We gaan ervan uit dat er in de toekomst nog een groter project zal komen waarin, gebruikmakend van de procedure, ook de tussenliggende edities aan de dataset zullen worden toegevoegd. Tabel 1 toont de specificaties van de verschillende edities.

Voor de jaren 1997-2008 is er gebruikt gemaakt van de cd-roms met almanakdata (uit de collectie van de Koninklijke Bibliotheek in Den Haag).

Tabel 1. Overzicht papieren edities

Editie	Uitgever	Redactie	Bladzijden
1956	Van Garde Drukkerij	D.J. Kappenburg	1464
1971	Van Garde Drukkerij	N.B.	796
1985	Bohn Stafleu van Loghum	J.J. Hanssen	876
1992-1993	Bohn Stafleu van Loghum	A.M. Garritsen	1128
1998-1999	Bohn Stafleu van Loghum	A.M. Garritsen	1492
1997-2009	Bohn Stafleu van Loghum	A.M. Garritsen	CD-ROM

Digitalisering

De verwerking van de papieren edities besloeg twee stappen: 1) het inscannen van de almanakken; 2) het verwerken van de scans tot verwerkbare *data entries*. Het scannen van de papieren almanakken is gedaan door Blomsma Printmedia. De almanakken zijn machinaal gescand op 600 dpi en als PDF-vorm per jaargang geleverd. Deze resolutie maakt het mogelijk om de scans via OCR (optical character recognition) naar tekstbestanden om te zetten. De PDF-bestanden zijn gedeponeerd (huidige eigenaar van de kopieerrechten is Thorald Wevers).

[2] We bedanken de redacteur Alice Garritsen voor het ter beschikking stellen van enkele van deze edities en de behulpzaamheid in de toelichting op de werkwijze van de gegevensverzameling.

De PDF's van de almanakken zijn in twee stappen verwerkt. Er zijn eerst handmatig annotaties gemaakt om alle entries en tussenkopjes te markeren. Deze entries vormen de basis van de datastructuur. Op basis van de annotaties zijn de pagina's gesegmenteerd om daar tekst uit te lezen via de Tesseract OCR[3] engine. De ingelezen platte tekst is opgesplitst in datavelden op basis van deze geïdentificeerde alinea's en meta-identificatie van het soort data en de veldnamen (bv 'adres', 'naam van de organisatie' etc). Op basis van de posities van woorden en regels in de afbeeldingen was het vervolgens mogelijk om de structuur van de tekst uit te werken en de in te lezen.

1992-1993 is de enige jaargang was met een consistente markering van datavelden. In de eerdere versies zijn er voor elk mogelijk veld (bijvoorbeeld bestuur of voorzitter) teveel verschillende gebruikte omschrijvingen om dit effectief samen te vatten. In de eerdere versies hebben organisaties ogenschijnlijk meer vrijheid gehad om zelf te bepalen wat er in hun beschrijving stond. Zo zijn er in 1985 meer dan 32 verschillende beschrijvingen voor directies en/of directeuren. Als gevolg is de resulterende data ook minder gestructureerd dan de nieuwere datasets. De gegevens zijn wel volledig opgenomen in het uitgebreide json-bestand maar aanvullende meta-identificatie is nodig om alle gegevens op tussen organisaties en edities vergelijkbare wijze te structureren.

Het resultaat is afhankelijk van de kwaliteit van de scans. In 1954 en 1971 geeft een combinatie van ouderdom en slechtere druktechniek (minder precieze letters) een hoger percentage fouten. Het gevolg hiervan is dat het ontdekken van structuur ook moeilijker is. Doordat bij deze jaren de structurerende elementen (bijvoorbeeld " :" of ".") niet perfect zijn ingelezen, was de structuur die op deze manier uit de documenten werd gehaald minder goed bruikbaar voor de dataset. Voor de meeste toepassingen moet dit geen probleem zijn. Het kan echter wel voorkomen dat bijvoorbeeld een specifieke naam niet gevonden wordt terwijl deze wel in de almanak staat, maar door lagere druk- en zettechniek van deze edities niet herkend wordt in de juist spelling. De dataset richt zich op de inhoudelijke secties van de almanak die betrekking hebben op organisaties (de sectie over de adel en de koninklijke familie is bijvoorbeeld altijd overgeslagen).

Dataset

De dataset is zodanig opgezet dat de gegevens uit de papieren edities zoveel mogelijk ontsloten wordt. Met het oog op de vele mutaties die de almanakken door de jaren heen doormaken, evenals de uiteenlopende manieren van hoe de data gestructureerd is, is er besloten om de dataset op te zetten in het flexibele .json-bestandsformaat. De gegevens staan dan niet in een gewoon tabel met altijd dezelfde kolommen maar iedere eenheid in het bestand kan hierin verschillende informatie-

[3] https://opensource.google.com/projects/tesseract

onderdelen hebben. Dit is nodig omdat, ten eerste, een groot deel van de data is genest, wat betekent dat datavelden meerdere elementen kunnen bevatten. Zo hebben besturen verschillende aantallen leden en bieden de bestanden per bestuurslid dan ook verschillende variabelen. Voor de jaargangen voorafgaand aan 1997 geldt dat er bijvoorbeeld ruimte was voor het vermelden van verschillende soorten besturen of commissies binnen organisaties. Ten tweede, staat er bij iedere 'record' in de almanak-editie, zeker bij de oudere edities, verschillende soorten opmerkingen. Er is dus veel variatie aan velden en waardes. Om het mogelijk te maken om deze data te behouden zouden er veel categorieën moeten worden samengevoegd of gecreëerd. Dit is veel werk doordat alle informatie inhoudelijk (door mensen) herkend en gematcht moeten worden. Verder en belangrijker is dat er ook keuzes gemaakt zouden moeten worden over de indeling van de informatie en een deel van de informatie wordt daarmee al gefilterd en geïnterpreteerd. Dit zou de toepassingsreikwijdte kunnen versmallen. Door het .json-formaat te gebruiken blijft de informatie zoveel mogelijk in de originele vorm (zoals het in de betreffende almanak staat) bewaard en kan het op die manier wijd toepasbaar blijven zonder data te verliezen door keuzes over de structurering. Aangezien het gebruik van de dataset niet vast ligt, is een zo wijd mogelijke toepasselijkheid gekozen.

JavaScript Object Notation, afgekort .json, is een data-uitwisselbestandstype waarin de data genest kan worden, maar nog relatief makkelijk te gebruiken is. Zo kunnen er *key-value pairs* gemaakt worden binnen een .json-bestand, met de mogelijkheid om daarin ook weer objecten te *embedden* (zie figuur 1). Dit sluit goed aan bij de geneste vorm van informatie in de almanakken, almede de vele verschillende waardes die de variabelen kunnen bevatten. De data die gewenst is, kan vervolgens hieraan worden onttrokken. Met het onderstaande codeboek zijn de jaargangen met meer structuur in verschillende programmeertalen gemakkelijk doorzoekbaar.

Er zijn twee CSV-bestanden toegevoegd. Vanaf 1997 (97-08.CSV) zijn de oorspronkelijke velden ver genoeg gestandaardiseerd om ze in een tabel te weergeven. Bij de eerdere versies is er een selectie van variabelen opgenomen in het CSV (zie Codeboek).

Tabel 2. Databronnen

Format	Bron	Jaargangen
PDF / OCR platte tekst	Boek, scan 600dpi	1954, 1971, 1985, 1992-1993, 1998-1999[4]
C-data (*custom format*)	CD	1997-1998, 1998, 1998-1999, 1999, 2002-2003, 2003, 2003-2004, 2004
HTML	CD	2004-2005, 2005, 2005-2006, 2006-2007, 2007, 2008, 2008-2009, 2009

De eerste dataset (97-08.json) is gebaseerd op de cd-roms. Bij het inlezen van de data deden zich enkele problemen voor als gevolg van de veranderde structuur en aanpak van de uitgever met betrekking tot de data. Tabel 2 laat zien welk format de data oorspronkelijk had. Tot 2004 werkten deze cd's met een eigen dataformat, gelijkend op C-data (*character data)*. Na 2004 is de uitgever overgestapt op HTML (*hardcoded*) als dataformat. De latere jaargangen waren overzichtelijk in HTML gestructureerd, en eenvoudig in te inlezen.[5]

De C-data van de eerdere cd-roms konden niet meteen eenvoudig samengevoegd worden. De cd-roms hadden installatieprogramma's die slechts te draaien waren op 32-bit systemen (*16-bit installer*). De almanakdata stond in een bestanden die opgebouwd waren aan de hand van *markers* (<!). Deze *markers* gaven aan dat een dataveld begon met daarbij een bijbehorend karakter om aan te geven wat er in dat dataveld stond. De structuur van de data kon vervolgend worden achterhaald door gebruik te maken van een ander bestand dat de veldnamen bevatte op volgorde van de bijbehorende integerwaarde van de karakters die velden markeerden. De C-data en HTML formats van de CD-ROM's hebben hierdoor uiteindelijk een overzichtelijke structuur met vergelijkbare velden.[6] Na het inlezen zijn de datasets vanaf 1997 samengevoegd, waarbij de organisaties over de jaargangen aan elkaar gekoppeld zijn.

[4] 1998-1999 is in het meest werkbare format verwerkt
[5] Hierbij is er gebruik gemaakt van BeautifulSoup voor Python, met lxml als *backend*
[6] De gegevens op de C-data CD-ROMS staan in een bestand dat is gestructureerd aan de hand van markers (<!) die aangeven dat een dataveld begint met daarbij een bijbehorend karakter om aan te geven wat er in dat dataveld stond. Aan de hand van een ander bestand (dit bestand bevatte veldnamen op volgorde van de bijbehorende integerwaarde van de karakters die velden markeerden) is de structuur bepaald.

Bij het verbinden van de jaren zijn we terughoudend geweest met samenvoegingen in gevallen waarbij het niet volledig zeker is dat het om dezelfde organisaties in verschillende jaargangen gaat. Twee punten van aandacht waren verhuizingen en naamswijzigingen. Bij verhuizingen zijn de organisaties goed te traceren wanneer een organisatie slechts van adres wijzigt. In elke jaargang staan er een aantal identieke organisatienamen en hierdoor moesten verhuizingen soms handmatig geïdentificeerd worden. Bij; 2) naamswijzigingen: hierbij verandert de naam en het adres niet, maar blijft de naam hetzelfde. Vaak delen veel organisaties hetzelfde adres, zeker wanneer het afdelingen van verenigingen betreft. Dit kan als gevolg hebben dat er ten onrechte organisaties aan elkaar worden gekoppeld. Daarom is er een maatstaf gehanteerd voor gelijkenis in namen en kozen we de naam met de grootste overlap boven een drempelwaarde kwam.[7] Dit bleek erg effectief, voor zowel naamswijzigingen als verschillende notaties voor namen. Fusies zijn buiten beschouwing gelaten.

De tweede dataset (boeken.json) is gebaseerd op de papieren versie. De dataset is zodanig opgezet dat de informatie van de Almanakken zo veel mogelijk te behouden blijft. Hierdoor is er veel informatie per organisatie die relatief ongestructureerd is (ie zonder meta-informatie die directe vergelijking met andere organisaties mogelijk maakt) en als tekst is toegevoegd. Met het bijgeleverde codeboek zijn de jaargangen met meer structuur in verschillende programmeertalen makkelijk doorzoekbaar. De oorspronkelijke inhoudelijke categorisering is ook behouden gebleven en meegeleverd in aparte TXT-bestanden per editie (editiejaar.TXT).

Gebruik

Voor het gebruik van de gegevens uit de JSON bestanen kan er met behulp van verschillende programma's, zoals Python, gericht gegevens onttrokken worden. Hieronder staat beschreven hoe dit in Python (zie Python.org) werkt.

Inlezen

```
1.  import json
2.
3.  fname = 'pyttersen.json' #Bestandsnaam of pad van databestand
4.
5.  with open(fname, encoding = 'utf-8') as f:
6.      data = json.load(f)
```

[7] Geoperationaliseerd met de cosine similarity om de rol van lengte van de naam een kleinere rol te laten spelen. Op deze manier matchen lange namen niet disproportioneel sneller, simpelweg omdat ze meer verschillende woorden hebben. De mate van overeenkomst wordt uitgedrukt met een getal tussen 0 en 1.

Voorbeeld: Personen onttrekken

In Python heeft de dataset het DICT dataformat, een object dat velden heeft met namen. De velden zijn aan te roepen met [], bijvoorbeeld via ORGANISATIE['NAAM']. Deze objecten zijn binnen Python volledig geoptimaliseerd om over te itereren. Om te ontdekken welke velden een dictionary heeft bestaat de methode .keys(), om te itereren over een *dictionary* (*keys* en *data*) bestaat de methode .items().

De structuur van de data is als volgt:

- De *dictionary* bevat organisaties met een unieke id. Hierbinnen bestaat een entry per versie waar de organisatie in voorkomt. Daarbinnen staat de *entry* van dat jaar volgens velden in het codeboek.

```
#Aanname hier is dat het bestand volgens bovenstaande is ingelezen
from csv import DictWriter

out_file = 'personen.csv'

fieldnames = ['organisatie','jaar','orgaan','aanhef', 'achternaam',
              'achtervoegsel', 'email', 'functie', 'initialen',
              'overig', 'politieke partij', 'titel',
              'toevoeging', 'website','text']

with open(out_file, 'w+', encoding = 'utf-8') as f:
    writer = DictWriter(f, fieldnames = fieldnames, lineterminator = '\n')
    writer.writeheader()
    for _,organisatie in data.items():
        for version, entry in organisatie.items():
            if 'fungroepen' in entry:
                for orgaan, personen in entry['fungroepen'].items():
                    for persoon in personen:
                        if isinstance(persoon, dict):
                            persoondata = {'organisatie':entry['naam'],
                                           'jaar': version,
                                           'orgaan':orgaan}
                            persoondata.update(persoon)
                            writer.writerow(persoondata)
```

Codebook JSON

Dit codebook legt de structuur uit van de json data van Pyttersen's almanak, edities 1997 tot 2009. Het is opgebouwd in twee losse onderdelen, één voor de data van 1997 tot 2009 (digitale bronnen) en één voor de data uit boeken (1954, 1971, 1985 en 1992-1993).

1997-2009

De jaargangen 1997-2009 hadden een CD bijgevoegd bij de boeken. Van deze CD's hebben we de gegevens uitgelezen. Deze gegevens waren redelijk goed gestructureerd. Per dataformat wisselden enkele veldnamen of waren sommige gegevens wel of niet opgenomen. Deze dataset bevat de gegevens zo volledig mogelijk. Sommige velden bestonden slechts in sommige jaargangen, die zijn dan wel opgenomen.

Beschrijving

De volgende edities zijn opgenomen:

Editie	Aantal organisaties
1997	19866
1998	17663
1997-1998	17582
1999	16412
1998-1999	16352
2002-2003	15457
2004	14739
2003-2004	14657
2003	14652
2004-2005	14623
2005	14564
2005-2006	14554
2006-2007	14352
2007	14142
2008	14068
2008-2009	13774
2009	13739

Door de jaren heen is de akmanak op vaak tweemaal per jaar uitgebracht. Bovenstaande edities hebben wij digitale edities van in ons bezit kunnen krijgen. Indien er een jaargang ontbreekt betekent dat niet dat die almanak niet is uitgekomen. Zoals ook is te zien in de aantallen organisaties zijn de verschillen tussen jaarlijkse en halfjaarlijkse edities minimaal.

Toelichting structuur

Deze dataset is opgebouwd als json, wat betekent dat het is opgedeeld in key-value paren. Per organisatie die voorkomt is er een unieke key, waarachter per jaargang van de almanak een nieuwe entry bestaat. De structuur van deze entries heeft varieert vanwege de structuur van de almanak. In principe bevat elk dataveld binnen een entry een tekst of string. Hierop bestaan uitzonderingen, namelijk:

- Als een organisatie meerdere keren voorkwam in één jaargang. Dit komt redelijk vaak voor, overal waar organisaties aan meerdere categorieën zijn toegekend. In dat geval bevatten datavelden een tekst als elke entry er dezelfde data voor had of een dictionary-format data met per unieke entry de data uit dat veld. Op die manier kunnen de oude entries gereconstrueerd worden. De daar gebruikte keys zijn uniek binnen de jaargang, maar omdat de entries zijn samengevoegd verwijzen ze niet meer naar bestaande entries.
- Enkele velden bevatten data van een andere vorm (list). Dit zijn de velden ‘nevenorganisaties’ en ‘tree adres (list)’.

Variabelen

c	**Betekenis**	**Toelichting**
naam	Naam van de organisatie	
toenaam	Eventuele toenaam, vaak een uitdrukking van de functie van de organisatie	
afkorting		
id	Unieke identifier van de organisatie	
versie	Versie van de Pyttersen Almanak.	Uitgedrukt als jaargang met jaartal(len)

tree adres	Pad van de organisatie binnen de structuur van de Pyttersen Almanak.	In de software van de Pyttersen Almanak zijn de categorieën geïmplementeerd in de vorm van een boom. Eerst splitst het op in in brede categorieën die daarna steeds specifieker worden. Het adres laat zien welke takken gevolgd moeten worden om tot de organisatie te komen. Bijvoorbeeld: Organisaties, stichtingen en verenigingen > Levensbeschouwing > Levensbeschouwelijke groeperingen > Overige. Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
tree adres (list)	Hetzelfde als bovenstaande, maar dan met de verschillende elementen opgesplitst en als list bewaard.	
Categorie	Hoofdcategorie.	Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Categorie - code	Code die gebruikt wordt voor de hoofdcategorie.	Codes kunnen veranderen per jaargang. Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Subcategorie 1 - code	Code die gebruikt wordt voor de eerste subcategorie. Codes kunnen veranderen per jaargang.	Codes kunnen veranderen per jaargang. Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Subcategorie 1 - naam	Eerste subcategorie	Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.

Subcategorie 2 - code	Code die gebruikt wordt voor de tweede subcategorie. Codes kunnen veranderen per jaargang.	Codes kunnen veranderen per jaargang. Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Subcategorie 2 - naam	Tweede subcategorie	Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Subcategorie 3 - code	Code die gebruikt wordt voor de derde subcategorie. Codes kunnen veranderen per jaargang.	Codes kunnen veranderen per jaargang. Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Subcategorie 3 - naam	Derde subcategorie	Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Subcategorie 4 - code	Code die gebruikt wordt voor de vierde subcategorie. Codes kunnen veranderen per jaargang.	Codes kunnen veranderen per jaargang. Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Subcategorie 4 - naam	Vierde subcategorie	Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
Subcategorie extra - code	Extra subcategorie, niet benoemd in de software, maar de code bestaat wel.	Codes kunnen veranderen per jaargang. Indien de organisatie in meer categorieën voorkomt zijn die geschikt in een dictionary-format met unieke keys per entry.
is_nevenorg	TRUE als de organisatie geregistreerd is als nevenorganisatie.	

org_type	Extra informatie over het type organisatie	Beschikbaar vanaf 2004-2005
neven_aanduiding	Aanduiding van nevenorganisaties van deze organisatie	
nevenorganisaties	Identifiers van de nevenorganisaties	Vrijwel alle nevenorganisaties zijn ook zelfstandig opgenomen in de Almanak. Als ze zijn opgenomen bevat deze list de identifier van de organisatie. Indien niet (hoofdzakelijk het geval in de oudste versies), dan bevat het de gegevens van de nevenorganisatie als dictionary.
hoofd_org	Identifier van de organisatie waarvan deze een nevenorganisatie is.	
bezoekadres - adres		
bezoekadres - plaats		
bezoekadres - postcode		
bezoekadres - postcode2	Opgenomen indien er een extra postcode was opgegeven.	In sommige gevallen was de postcode enkel gegeven in het postcode2-veld. In dat geval bestaat er geen waarde onder 'bezoekaders – postcode', maar wel onder 'bezoekadres – postcode2'
postadres - adres		
postadres - plaats		
postadres - postcode		
postadres - postcode2	Opgenomen indien er een extra postcode was opgegeven	In sommige gevallen was de postcode enkel gegeven in het postcode2-veld. In dat geval bestaat er geen waarde onder 'postadres – postcode', maar wel onder 'postadres – postcode2'
provincie		
telefoon		
fax		
email		
www		
doelstelling		
overige info		
in- & ext. betr.		
inlichtingen		
instellingen		

aantal inwoners		Alleen bij gemeente of provincie
afdelingen		
archief		Alleen bij gemeente
arrondissement		Alleen bij gemeente
kantongerecht		Alleen bij gemeente
land		Bij organisaties die gevestigd zijn in het buitenland, bijv. Ambassades
nevenvestiging		
openingstijden		
opleidingen		
oppervlakte		Alleen bij gemeente of provincie
oprichtingsdatum		
orgaan	Tijdschrift of andere communicatie die de organisatie uitbrengt	
politie	Politieregio	Alleen bij gemeente
samenwerk. org.	Regionale samenwerkingsorganisatie	Alleen bij gemeente
steden en dorpen		Alleen bij gemeente
bibliotheek		Alleen bij gemeente
bureau decanen		
studentendecanen		
transferpunt		
vakgroepen		
vestiging(en)		
wetensch. winkel		
collectie		Bij musea
zwaartepunt collectie		Bij musea
EXTBOOMID		Bestaat slechts in de edities 2004-2005 tot en met 2007. Deze werkte met een duidelijke boomstructuur, meegegeven in de entries.
fungroepen	Zie toelichting over fungroepen	
shortName		

nodeId		Bestaat slechts in de edities 2004-2005 tot en met 2007. Deze werkte met een duidelijke boomstructuur, meegegeven in de entries.
tocOrd		
POPROV	Provincie voor het postadres, indien van toepassing	
boomTak		Bestaat slechts in de edities 2004-2005 tot en met 2007. Deze werkte met een duidelijke boomstructuur, meegegeven in de entries.
boomOrd		Bestaat slechts in de edities 2004-2005 tot en met 2007. Deze werkte met een duidelijke boomstructuur, meegegeven in de entries.
boomId	Dit is de unieke identifier die het pad door de boom volgt. De hoofd-organisaties van nevenorganisaties kunnen geïdentificeerd worden door één stap terug te gaan in de boom hier. Dit werkt minder goed in 2005, waar het voor sommige entries inconsistent is.	Bestaat slechts in de edities 2004-2005 tot en met 2007. Deze werkte met een duidelijke boomstructuur, meegegeven in de entries.
___Nr		Bestaat slechts in de edities 2004-2005 tot en met 2007. Deze werkte met een duidelijke boomstructuur, meegegeven in de entries.
onam	Organisatienaam	
codegetal	Dit getal bestond in de versies tot en met 2004 en is een enigszins consistente identifier tussen die versies. De waarden zijn echter niet altijd uniek.	
zoekId		Bestaat slechts in de edities 2004-2005 tot en met 2007. Deze werkte met een duidelijke boomstructuur, meegegeven in de entries.
OID	Organisatie id, uniek en consistent binnen jaargangen. Deze identifiers zijn gebruikt om samengevoegde organisaties te structureren.	

daOrigFNm		Bestaat slechts in de edities 2004-2005 tot en met 2007. Deze werkte met een duidelijke boomstructuur, meegegeven in de entries.

Fungroepen (bestuursonderdelen)

De verschillende almanakken bevatten verschillende groepen betrokken personen (in de digitale structuur van de Almanakken aangeduid als fungroepen). Om deze bij elkaar te houden hebben we ze allemaal achter de key 'fungroepen' bewaard. Hierachter bestaat een aparte datastructuur met de volgende opbouw:

1. Per bestuursonderdeel een key, de key is de naam van het bestuursonderdeel.
2. Binnen het bestuursonderdeel een lijst van de personen.
3. De personen zijn intern opgedeeld met de volgende keys:

Veld	**Toelichting**
aanhef	
achternaam	
achtervoegsel	Titels die achter de naam worden geplaats, bijvoorbeeld MSc, PhD
email	
functie	
initialen	
overig	
politieke partij	
titel	
toevoeging	
website	
text	Waar inlezen niet werkte is gekozen de entries hier op te slaan. Dit gaat om ongeveer 0.03% van de cases.

De volgende groepen bestaan binnen de fungroepen:

Algemeen secretaris	LPF
Ambtelijk secretaris	Leden
Bestuur	Medisch werkzame personen
Burgemeester	OSF
CDA	President
ChristenUnie	PvdA
ChristenUnie-SGP	Raad van Advies
College van Bestuur	Raad van Bestuur
Commissaris van de Koningin	Raad van Commissarissen
D66	Raad van Toezicht
Directie	SP
Europa Transparant	Secretaris
Fractievoorzitters	VVD
Gedeputeerden	Vicevoorzitter
Gemeentesecretaris	Voorzitter
Griffier	Wethouders
GroenLinks	bestuur/directie
Hoofd Communicatie	conservator
Hoofd Voorlichting	minister
Hoofdadvocaat-Generaal	proc.-generaal
Hoofdofficier van Justitie	secr.-generaal
Korpsbeheerder	staatssecr.
Korpschef	univ. raad

Boeken (1956, 1971, 1985, 1992-1993)

In tegenstelling tot de digitale bronnen zijn de boeken relatief ongestructureerd. Er lijkt meer vrijheid te bestaan voor de verschillende velden en de formats voor het invoegen van besturen. Dat, gekoppeld met onzekerheden uit de OCR (optical character recognition; het automatisch inlezen van tekst) maakt dat de data minder gestructureerd is dan de data uit digitale bronnen.

Ten eerste bevatten 1956 en 1971 veel onzekerheden. Dat betekent dat het uitlezen van een volledige structuur de mogelijkheid tot fouten openzet met als gevolg een opeenstapeling van fouten. Daarom hebben we de dataset met weinig bewerkingen ingevoegd. De titel, hoofdstukken

en categorieën zijn toegevoegd en de tekst is gestructureerd in de paragrafen die werden geïdentificeerd door Tesseract OCR. Deze bleek in de latere edities de verschillende velden goed uit elkaar te halen.

1985 en 1992-1993 zijn opgesplitst naar de velden met identifier voor die velden omdat die structuur in de boeken beter tot uiting kwam. De velden zijn echter niet in eenzelfde mate gestandaardiseerd als in de digitale edities. In 1985 bijvoorbeeld hebben wij 420 verschillende velden geïdentificeerd, waarbij we ook verschillende spellingen als verschillende velden tellen. Als gevolg is de data minder goed te beschrijven zoals hierboven. De veldnamen spreken echter doorgaans voor zich.

Voor 1992-1993 was de kwaliteit van de OCR en de standaardisering van de velden zodanig goed dat die editie op dezelfde manier is verwerkt als de digitale bronnen. Met name de personen zijn nagenoeg in dezelfde kwaliteit als de digitale bronnen, met als enige kanttekening dat de OCR inleesfouten oplevert. Dit uit zich voornamelijk in missende letters.

Datastructuur

Net als de dataset samengesteld uit digitale bronnen is deze opgebouwd met een unieke key per organisatie. Daarbinnen volgt per jaargang een key. Echter, in tegenstelling tot de andere dataset zijn de organisaties hier tussen de versies niet gematcht. We bieden wel suggesties tot matchen, zie daarvoor de header ‘matching’.

Velden

Binnen elke entry is een aantal velden mogelijk, parallel aan de andere versie

Variabele	Betekenis	Toelichting
naam	Naam van de organisatie	
OID	Unieke identifier van de organisatie binnen de jaargang	
versie	Versie van de Pyttersen Almanak.	
tree adres	Pad van de organisatie binnen de structuur van de Pyttersen Almanak.	1992-1993: Indien er meerdere versies voorkwamen staan deze in een dictionary met de OIDs als keys.
page	Paginanummer. NB: Dit is het paginanummer in de tevens op DANS beschikbaar gemaakte pdf’s van de boeken.	1992-1993: Indien er meerdere versies voorkwamen staan deze in een dictionary met de OIDs als keys.
Categorie	Hoofdcategorie.	1992-1993: Indien er meerdere versies

		voorkwamen staan deze in een dictionary met de OIDs als keys.
Subcategorie 1 – naam	Eerste subcategorie	1992-1993: Indien er meerdere versies voorkwamen staan deze in een dictionary met de OIDs als keys.
Subcategorie 2 – naam	Tweede subcategorie	1992-1993: Indien er meerdere versies voorkwamen staan deze in een dictionary met de OIDs als keys.
Subcategorie 3 – naam	Derde subcategorie	1992-1993: Indien er meerdere versies voorkwamen staan deze in een dictionary met de OIDs als keys.
bezoekadres – straat		1992-1993
bezoekadres – nummer		1992-1993
postbus		1992-1993
postcode		1992-1993
plaatsnaam		1992-1993
telefoon		1992-1993
fax		1992-1993
text	Dit veld bevat alle tekst, opgesplitst in paragrafen die niet automatisch aan een veld toegewezen kon worden. Dit bestaat in de versies 1956, 1971 en 1985. Voor de versie 1992-1993 hebben wij de restanten handmatig gecorrigeerd.	
Matches	Per andere jaargang de beste matches (maximaal 5, mits deze een *cosine similarity* hadden van .6)	Zie kopje Matching

Bestuursonderdelen (fungroepen,; 1992-1993)

In 1992-1993 was de kwaliteit van de OCR zodanig goed dat we geheel consistent met de digitale bronnen de verschilende bestuursonderdelen hebben geïdentificeerd en de betrokken personen hebben gestructureerd volgens dezelfde logica. De datastructuur is identiek, dus elke analyse op personen voor de digitale bronnen zal hier ook werken.

Matching

Zoals boven aangegeven hebben wij in deze dataset geen organisaties gematcht tussen de jaargangen. De reden dat we dit niet hebben gedaan was dat we in de nieuwere data voor elke jaargang een adres hadden om de match voor te controleren. Voor 1992-1993 was de toevoeging van een adres eerder uitzondering dan regel. We hebben wel dezelfde berekening uitgevoerd, gebruik makende van de *cosine similarity*[8]. Dit resulteert in een aantal voorgestelde matches die de onderzoeker naar eigen inzicht kan gebruiken.

Als vuistregel voor de matching willen we meegeven dat vooral korte organisatienamen met veel generieke woorden snel een sterke overeenkomst met een organisatienaam met dezelfde generieke woorden aangeeft. Zo wordt 'De Nationale Padvindstersraad' makkelijk gematcht met 'De nationale Investeringsbank (N.V.)'. Door het kiezen van een drempelwaarde en bijvoorbeeld een controle of de organisaties wederzijds de beste match zijn kan deze maatstaf betrouwbaarder worden ingezet.

Gepaard met de check op postcode bleek een drempelwaarde van 0.6 goed te presteren in de data uit digitale bronnen. Hier zijn voor elke jaargang vijf opties meegegeven, mits deze boven de drempelwaarde van 0.3 uit kwamen. 0.3 is een extreem lage overeenkomstscore. De drempelwaarde dient om de dataset niet op te blazen om voorstellen te doen voor organisaties waarvoor met deze methode echt geen match gevonden kan worden.

Onderstaande geeft een voorbeeld van deze structuur voor de 'Vereniging van Oud-militairen van het Koninklijk Nederlands-Indische Leger "MADJOE"' (1985). De onderste drie matches zijn steeds weggelaten voor de overzichtelijkheid. Elke entry bevat de naam van de voorgestelde match, de ID van deze organisatie en de bijbehorende *cosine similarity*.

[8] De *cosine similarity* maatstaf vergelijkt documenten, opgesteld als vectoren van het aantal keren dat woorden voorkomen. Hierbij telt de lengte van het document minder mee dan bij bijvoorbeeld de Levenshtein distance, die het aantal verschuivingen telt om van het ene document naar het andere te gaan. De *cosine similarity* heeft hier als nadeel dat verschillend gespelde of verschillend ingelezen woorden als andere woorden worden geteld.

```
"Matches": {
        "1956": [
          ["\u201eVereniging van Vrienden van het \u201a Goo\u00efreservaat\u2019\u2019.", 776,
            0.6396021490668312],
          ["Vereniging van Hoofdambtenaren van __ het Kadaster,", 1048, 0.6396021490668312],
          (…)
        ],
        "1971": [
          ["Vereniging van Oud-Onderofficieren van het Koninklijk Nedertands-Indische Leger
\u201eMADJOE\u201d.", 2890, 0.8181818181818182],
          ["\u201aVereniging van Vrienden van het Gooi\u201d.", 3426, 0.6396021490668312],
          (…)
        ],
        "1992-1993": [
          ["Vereniging van Oud-Militairen van het Koninklijk Nederlands-Indisch Leger Madjoe", 10633,
            0.9090909090909092],
          ["Vereniging van Vrienden van het Gooi", 10848, 0.6396021490668312],
          (…)
        ]
      }
```

Codeboek CSV-bestanden

Dit codebook legt de structuur uit van de CSV data van Pyttersen's almanak, edities 1997 tot 2008. Aan het einde volgt nog een korte toelichting voor het CSV-bestand bij de boeken (1956, 1971, 1985, 1992-1993). Het is gestructureerd in drie onderdelen. Eerst beschrijven we kort de data. Vervolgens bespreken we in twee onderdelen de variabelen, waarin we eerst de variabelen die de organisatie beschrijven, waarbij we waar nodig de betekenis toelichten. Vervolgens besteden we kort aandacht aan de structurering van gegevens van de verschillende bestuursonderdelen. In de bijgevoegde .json-versie zijn de personen verder opgesplitst te vinden met aparte velden voor initialen, achternaam etc. Om dit in tabelvorm te gieten hebben we deze teruggebracht naar één standaardformat dat verderop beschreven is.

Beschrijving

De volgende edities zijn opgenomen:

Editie	Aantal organisaties
1997	19866
1998	17663
1997-1998	17582
1999	16412
1998-1999	16352
2002-2003	15457
2004	14739
2003-2004	14657
2003	14652
2004-2005	14623
2005	14564
2005-2006	14554
2006-2007	14352
2007	14142
2008	14068
2008-2009	13774
2009	13739

Door de jaren heen is de akmanak op vaak tweemaal per jaar uitgebracht. Bovenstaande edities hebben wij digitale edities van in ons bezit kunnen krijgen. Indien er een jaargang ontbreekt betekent dat niet dat die almanak niet is uitgekomen. Zoals ook is te zien in de aantallen organisaties zijn de verschillen tussen jaarlijkse en halfjaarlijkse edities minimaal.

Variabelen

Variabele	Betekenis	Toelichting

naam	Naam van de organisatie	
toenaam	Eventuele toenaam, vaak een uitdrukking van de functie van de organisatie	
afkorting		
id	Unieke identifier van de organisatie	
versie	Versie van de Pyttersen Almanak.	Uitgedrukt als jaargang met jaartal(len)
tree adres	Pad van de organisatie binnen de structuur van de Pyttersen Almanak.	In de software van de Pyttersen Almanak zijn de categorieën geïmplementeerd in de vorm van een boom. Eerst splitst het op in in brede categorieën die daarna steeds specifieker worden. Het adres laat zien welke takken gevolgd moeten worden om tot de organisatie te komen. Bijvoorbeeld: Organisaties, stichtingen en verenigingen > Levensbeschouwing > Levensbeschouwelijke groeperingen > Overige. Indien de organisatie in meer categorieën voorkomt zijn die gescheiden door een puntkomma (;).
Categorie	Hoofdcategorie.	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Categorie - code	Code die gebruikt wordt voor de hoofdcategorie.	Codes kunnen veranderen per jaargang. Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 1 - code	Code die gebruikt wordt voor de eerste subcategorie. Codes kunnen veranderen per jaargang.	Codes kunnen veranderen per jaargang. Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 1 - naam	Eerste subcategorie	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 2 - code	Code die gebruikt wordt voor de tweede subcategorie. Codes kunnen veranderen per jaargang.	Codes kunnen veranderen per jaargang. Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 2 - naam	Tweede subcategorie	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.

Subcategorie 3 - code	Code die gebruikt wordt voor de derde subcategorie. Codes kunnen veranderen per jaargang.	Codes kunnen veranderen per jaargang. Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 3 - naam	Derde subcategorie	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 4 - code	Code die gebruikt wordt voor de vierde subcategorie. Codes kunnen veranderen per jaargang.	Codes kunnen veranderen per jaargang. Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 4 - naam	Vierde subcategorie	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie extra - code	Extra subcategorie, niet benoemd in de software, maar de code bestaat wel.	Codes kunnen veranderen per jaargang. Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
is_nevenorg	TRUE als de organisatie geregistreerd is als nevenorganisatie.	
org_type	Extra informatie over het type organisatie	Beschikbaar vanaf 2004-2005
neven_aanduiding	Aanduiding van nevenorganisaties van deze organisatie	
bezoekadres - adres		
bezoekadres - plaats		
bezoekadres - postcode		
bezoekadres - postcode2		
postadres - adres		
postadres - plaats		
postadres - postcode		
provincie		
telefoon		
fax		
email		
www		
doelstelling		

overige info		
in- & ext. betr.		
inlichtingen		
instellingen		
aantal inwoners		Alleen bij gemeente of provincie
afdelingen		
archief		Alleen bij gemeente
arrondissement		Alleen bij gemeente
kantongerecht		Alleen bij gemeente
land		Bij organisaties die gevestigd zijn in het buitenland, bijv. Ambassades
nevenvestiging		
openingstijden		
opleidingen		
oppervlakte		Alleen bij gemeente of provincie
oprichtingsdatum		
orgaan	Tijdschrift of andere communicatie die de organisatie uitbrengt	
politie	Politieregio	Alleen bij gemeente
samenwerk. org.	Regionale samenwerkingsorgani-satie	Alleen bij gemeente
steden en dorpen		Alleen bij gemeente
bibliotheek		Alleen bij gemeente
bureau decanen		
studentendecane n		
transferpunt		
vakgroepen		
vestiging(en)		
wetensch. winkel		
collectie		Bij musea
zwaartepunt collectie		Bij musea

Bestuursonderdelen

De volgende onderdelen bevatten alle bestuurders. Personen worden binnen een orgaan gescheiden met puntkomma's (;) en zijn als volgt gestructureerd:

Naam

1. Aanhef (bijvoorbeeld dhr., mevr.)
2. Titels
3. Initialen
4. Achternaam
5. Achtervoegsel (bijvoorbeeld MBA, RA, RI)

Overige gegevens

In het geval van een niet gevulde positie zal hier 'vacant' staan. De volgende velden zijn dan gescheiden met komma's:

6. Functie
7. Emailadres. Komt slechts voor bij sommige politieke organen. (email: [adres])
8. Persoonlijke website. Komt slechts voor bij sommige politieke organen. (website: [website])

Indien de persoon een andere organisatie vertegenwoordigt in een bestuur staat dit tussen haakjes achter de functie of – indien er geen functie is opgegeven – achter de naam. Eveneens kunnen overige toelichtingen, zoals bijvoorbeeld a.i. (ad interim) in deze positie tussen haakjes staan. Deze hebben doorgaans betrekking op de functie en komen dus vrijwel uitsluitend voor waar ook een functie is opgegeven.

Voorbeelden

- dhr. prof. dr. ir. R. Huirne, directeur wetenschap
- A. Kuyvenhoven, dir. (a.i.)
- dhr. B. Belder, email: bastiaan.belder@europarl.europa.eu

Velden

Algemeen secretaris	LPF
Ambtelijk secretaris	Leden
Bestuur	Medisch werkzame personen
Burgemeester	OSF
CDA	President
ChristenUnie	PvdA
ChristenUnie-SGP	Raad van Advies
College van Bestuur	Raad van Bestuur
Commissaris van de Koningin	Raad van Commissarissen
D66	Raad van Toezicht

Directie	SP
Europa Transparant	Secretaris
Fractievoorzitters	VVD
Gedeputeerden	Vicevoorzitter
Gemeentesecretaris	Voorzitter
Griffier	Wethouders
GroenLinks	bestuur/directie
Hoofd Communicatie	conservator
Hoofd Voorlichting	minister
Hoofdadvocaat-Generaal	proc.-generaal
Hoofdofficier van Justitie	secr.-generaal
Korpsbeheerder	staatssecr.
Korpschef	univ. raad

Boeken

Voor de data uit boeken is een klein CSV-bestand toegevoegd. Dit dient meet als overzicht dan als dataset. In de digitale edities zijn de velden in grote mate gestandaardiseerd, voor de boeken geldt dit niet. Alleen voor 1985 zou dit bijvoorbeeld al ongeveer 400 kolommen opleveren, met als gevolg een extreem onoverzichtelijke dataset. Alle data is wel beschikbaar in json-format en de boeken zijn als ingescande pdf's te downloaden via DANS. In dit CSV-bestand kan per organisatie gevonden worden hoe deze gecategoriseerd is, met welke ID deze in de json te vinden is en op welke pagina van de pdf deze begint. Ze bevat de volgende velden:

Variabele	**Betekenis**	**Toelichting**
naam	Naam van de organisatie	
versie	Versie van de almanak waarin deze organisatie voorkomt	
id	Unieke identifier in het json-bestand	
page	Pagina in de bijgeleverde pdf	NB: Niet het paginanummer van het boek zelf, maar van de pdf.
tree adres	Pad van de organisatie binnen de structuur van de Pyttersen Almanak.	In de software van de Pyttersen Almanak zijn de categorieën geïmplementeerd in de vorm van een boom. Eerst splitst het op in brede categorieën die daarna steeds specifieker worden. In de nieuwere edities wordt dit steeds specifieker. Het adres laat zien welke takken gevolgd moeten worden om tot de organisatie te komen. Bijvoorbeeld: Organisaties, stichtingen en verenigingen > Levensbeschouwing > Levensbeschouwelijke groeperingen > Overige. Indien de organisatie in meer categorieën voorkomt zijn die gescheiden door een puntkomma (;).

Categorie	Hoofdcategorie.	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 1 - naam	Eerste subcategorie	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 2 - naam	Tweede subcategorie	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.
Subcategorie 3 - naam	Derde subcategorie	Indien de organisatie in meerdere categorieën voorkomt zijn die gescheiden met een puntkomma (;). De volgorde van de categorieën is in elk veld gelijk.