

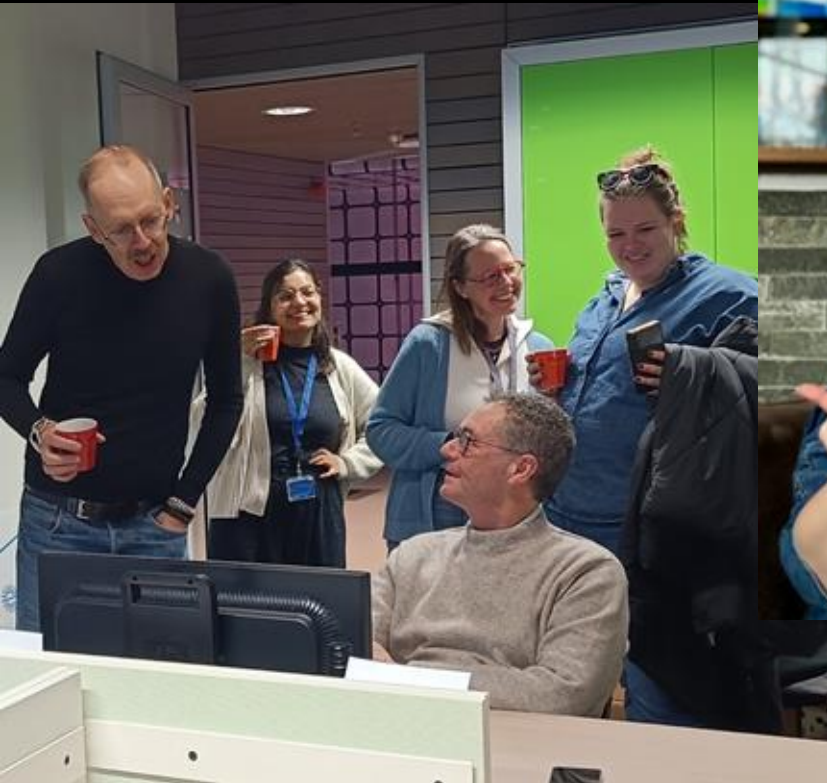


Nine Claassen
Netwerk Digitaal Erfgoed

**Shannon van
Muijden**
Rijksmuseum

2 datadudes
11 erfgoed datasets...

Uitdaging: zoveel mogelijk data schonen
in één dag



Aanpak

OpenRefine Stedelijk Museum Alkmaar thesaurus.xlsx Permalink

Facet / Filter Undo / Redo 0/0 < 2,910 rows Extensions Wikibase

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

	recordnummer	term_input	Column	Column2	aat_label (score)	Column3	aat_uri	match_score
1.	5198	gereedschap- en uitrustingsonderdelen			gereedschap en uitrustingsonderdel (0.99)		http://vocab.getty.edu/aat/300024871	0.99
2.	4155	vraag en antwoord spel			vraag en antwoordspel (0.98)		http://vocab.getty.edu/aat/300431615	0.98
3.	4065	horlogekettingsieraad			horlogekettingsieraad (0.98)		http://vocab.getty.edu/aat/300403914	0.98
4.	3552	Berlijns-zilver			berlijns zilver (0.97)		http://vocab.getty.edu/aat/300136730	0.97
5.	3137	kledingaccessoire			kledingaccessoires (0.97)		http://vocab.getty.edu/aat/300209273	0.97
6.	742	Procedés en Technieken			proceeds en techniek (0.97)		http://vocab.getty.edu/aat/300053001	0.97
7.	297	behendigheidspel			behendigheidsspel (0.97)		http://vocab.getty.edu/aat/300424546	0.97
8.	6533	Zuid-Nederland			zuidnederlands (0.96)		http://vocab.getty.edu/aat/300386537	0.96
9.	82	kunsthandel			kunsthandsel (0.96)		http://vocab.getty.edu/aat/300005232	0.96
10.	915	lagere school			lagere schol (0.96)		http://vocab.getty.edu/aat/300006507	0.96
11.	393	kleuren foto			kleurenfoto (0.96)		http://vocab.getty.edu/aat/300128359	0.96
12.	6058	medicijnpot			medicijnpott (0.96)		http://vocab.getty.edu/aat/300285260	0.96
13.	891	poppenmeubel			poppenmeubels (0.96)		http://vocab.getty.edu/aat/300423579	0.96
14.	4883	richting bord			richtingbord (0.96)		http://vocab.getty.edu/aat/300211858	0.96
15.	5160	compact disc			compact discs (0.96)		http://vocab.getty.edu/aat/300028673	0.96
16.	6044	compact disc			compact discs (0.96)		http://vocab.getty.edu/aat/300028673	0.96
17.	6540	rechts onder			rechtsonder (0.96)		http://vocab.getty.edu/aat/300404456	0.96
18.	1397	Vervoersmiddelen			vervoermiddel (0.96)		http://vocab.getty.edu/aat/300042929	0.96
19.	1032	objectgenre			objectgenres (0.96)		http://vocab.getty.edu/aat/300185711	0.96
20.	1036	bezinestation			benzinstation (0.96)		http://vocab.getty.edu/aat/300007815	0.96
21.	1961	kaasprikker			kaasprikker (0.96)		http://vocab.getty.edu/aat/300180554	0.96
22.	1389	picnickmand			picnickmand (0.96)		http://vocab.getty.edu/aat/300211978	0.96
23.	1480	koperslager			koperslagers (0.96)		http://vocab.getty.edu/aat/300025315	0.96
24.	2332	koperslager (7)			koperslagers (0.96)		http://vocab.getty.edu/aat/300025315	0.96
25.	4653	kaasdragers			kaasdragers (0.96)		http://vocab.getty.edu/aat/300411990	0.96
26.	5930	Massachusetts			massachusetts (0.96)		http://vocab.getty.edu/aat/300017558	0.96
27.	3417	patriotisme			patriotisme (0.96)		http://vocab.getty.edu/aat/300055531	0.96
28.	4808	straatsoeils			straatsoeiel (0.96)		http://vocab.getty.edu/aat/300010728	0.96

1393	Zeemeermin	objectnaam					https://data.cultureelerfgoed.nl/term/id/cht/27ac28e3-e69b-4bac
12	stoomfluit	objectnaam					https://data.cultureelerfgoed.nl/term/id/cht/292d741e-3a45-46ce
129	Informatieborden	objectcategorie		veiligheid	informatiebord		https://data.cultureelerfgoed.nl/term/id/cht/2e44391b-0677-4f76
459	informatiebord	objectcategorie	Informatie	promotiebord			https://data.cultureelerfgoed.nl/term/id/cht/2e44391b-0677-4f76
1366	eikenhout	materiaal					https://data.cultureelerfgoed.nl/term/id/cht/315c3975-51fa-4730
699	gashouders	objectnaam					https://data.cultureelerfgoed.nl/term/id/cht/3508834c-b08b-4d16
483	lino	objectnaam					https://data.cultureelerfgoed.nl/term/id/cht/439af2c2-e2c2-4e11
2189	moerasijzererts	objectnaam					https://data.cultureelerfgoed.nl/term/id/cht/446c5b35-c482-44c2
415	gemengde techniek	objectcategorie					https://data.cultureelerfgoed.nl/term/id/cht/49e38162-67c8-466e
478	ets E/D	techniek			zie term suggestie	etsen	https://data.cultureelerfgoed.nl/term/id/cht/4cb6455e-4c06-49f7

termennetwerk



DE JEUGD VANTEGE NWOOR DIG



- CHT en
s://data.
s://data.
s://data.
s://data.
s://data.
s://data.
s://data.
s://data.
s://data.
s://data.



datasetregister

s://data.cultureelerfgoed.nl/term/id/cht/1c08bf7-4f62-47f0-1
s://data.cultureelerfgoed.nl/term/id/cht/1c359a5c-19c2-42f1-
s://data.cultureelerfgoed.nl/term/id/cht/1fccff8c-eb79-48bf-a
s://data.cultureelerfgoed.nl/term/id/cht/20c6702f-15e7-4655
s://data.cultureelerfgoed.nl/term/id/cht/20c6702f-15e7-4655

Aanpak Nine opschonen

- Analyse in OpenRefine
 - Waar gaan de data over?
 - Wat zit er in de dataset?
 - Welke termen zijn dubbelop? Of synoniemen?
- Welke terminologiebronnen passen bij de data?
 - Matchen
 - Wat mist er (qua aanbod)
- Adviseren

Aanpak Shannon

1. Analyse in Excel

- Soorten termen (objectnamen / materialen etc.)
- Enkele woorden/zinnen

2. Bewerken in OpenRefine

- Gelijksoortige termen bij elkaar met cluster & edit
- Match met geschikte terminologiebron (AAT / CHT / Wikidata / WO2 thesaurus enz.)

3. Match met andere thesauri

- Match met thesaurus van Rijks maar ook eerder gematchede thesauri

4. Match met vector embeddings

A	B
term	term.soort
aan-/afmeren	activiteit
aanbrengen bodembesche	activiteit
aanbrengen steenglooing	activiteit
aankomst prins	activiteit
aamleggen dam	activiteit
aanleg bouwput	activiteit
aanleg jachthaven	activiteit
aanleg proefpolder	activiteit
aanleg weg	activiteit
aanleggen dam	activiteit
aanleggen haven	activiteit
aanleggen Haveneiland	activiteit
aanleggen proefpolder	activiteit

2,910 rows

Extensions Wikibase ▾

Show as: rows records Show: 5 10 25 50 100 500 1000 rows


« first < previous 1 - 10 next > last »


▼ All	▼ recordnummer	▼ term_input	▼ Column	▼ Column2	▼ aat_label (score)	▼ Column3	▼ aat_uri	▼ match_score
☆ ↗	1.	5198	gereedschap- en uitrustingsonderdelen		gereedschap en uitrustingsonderdel (0.99)		http://vocab.getty.edu/aat/300024871	0.99
☆ ↗	2.	4155	vraag en antwoord spel		vraag en antwoordspel (0.98)		http://vocab.getty.edu/aat/300431615	0.98
☆ ↗	3.	4065	horlogekettingsierraad		horlogekettingsieraad (0.98)		http://vocab.getty.edu/aat/300403914	0.98
☆ ↗	4.	3552	Berlijns-zilver		berlijns zilver (0.97)		http://vocab.getty.edu/aat/300136730	0.97
☆ ↗	5.	3137	kledingaccessoire		kledingaccessoires (0.97)		http://vocab.getty.edu/aat/300209273	0.97
☆ ↗	6.	742	Procédés en Technieken		proceeds en techniek (0.97)		http://vocab.getty.edu/aat/300053001	0.97
☆ ↗	7.	297	behandigheidspel		behandigheidsspel (0.97)		http://vocab.getty.edu/aat/300424546	0.97
☆ ↗	8.	6533	Zuid-Nederland		zuidnederlands (0.96)		http://vocab.getty.edu/aat/300386537	0.96
☆ ↗	9.	82	kunsthandel		kunsthandels (0.96)		http://vocab.getty.edu/aat/300005232	0.96
☆ ↗	10.	915	lagere school		lagere schol (0.96)		http://vocab.getty.edu/aat/300006597	0.96


	A	B	C	D	E	F	G
1	term	term.soort	term.statu	invoer.naam	opmerking term	suggestie term	URI
2	aan-/afmeren	activiteit			probeer de inhoudelijke informatie bij t	aan-/afmeren	
3	aanbrengen bodembesche	activiteit			probeer de inhoudelijke informatie bij t	aanbrengen	
4	aanbrengen steenglooing	activiteit			probeer de inhoudelijke informatie bij t	aanbrengen	
5	aankomst prins	activiteit			probeer de inhoudelijke informatie bij t	aankomen	
6	aamleggen dam	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
7	aanleg bouwput	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
8	aanleg jachthaven	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
9	aanleg proefpolder	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
10	aanleg weg	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
11	aanleggen dam	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
12	aanleggen haven	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
13	aanleggen Haveneiland	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
14	aanleggen proefpolder	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
15	aanleggen strekdam	activiteit			probeer de inhoudelijke informatie bij t	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec

Vector embedding van de AAT

record	A	B	C	D	E
numr	term_input	aat_label (score)	aat_uri	match	_score
205	schenking	schenking (1.00)	aat:300138913	100%	
3865	opdracht	opdracht (1.00)	aat:300026114	100%	
371	jacht	jacht (1.00)	aat:300082988	100%	
6424	varen	varen (1.00)	aat:300239493	100%	
377	veehoeder	veehoeder (1.00)	aat:300025615	100%	
370	vissen	vissen (1.00)	aat:30016988	100%	
7117	Bruinvis	bruinvis (1.00)	aat:300310627	100%	
7245	catalogusnummer	catalogusnummer (1.00)	aat:300004620	100%	
6574	inventarisnummer	inventarisnummer (1.00)	aat:3000312355	100%	
148	fantasie	fantasie (1.00)	aat:300068545	100%	
147	hond	hond (1.00)	aat:300310627	100%	
5229	constructie	constructie (1.00)	aat:300047195	100%	
4466	vierkant	vierkant (1.00)	aat:300055637	100%	
764	bedrijf	bedrijf (1.00)	aat:300000000	100%	
244	echtgenoot	echtgenoot (1.00)	aat:300154341	100%	
247	echtgenote	echtgenote (1.00)	aat:300154343	100%	
233	eigenaar	eigenaar (1.00)	aat:300203630	100%	
3071	gids	gids (1.00)	aat:300026300	100%	

Click the  icon to view the hierarchy.
Check boxes to view multiple records at once.

1.  **husbands**
(spouses, <people by family relationship>)

Click the  icon to view the hierarchy.
Check boxes to view multiple records at once.

1.  **wives**
(spouses, <people by family relationship>)

```
aat = pd.read_csv(aat_csv, sep=';')
```

```
aat_list = []
```

```
aat_uri = []
```

Conclusie

- ★ Het opschonen van data is goed te doen met de juiste tools
- ★ Deze kennis is goed over te dragen
- ★ Werken loont



Coming soon

Data-clean-festival

Houd de NDE-kanalen in de gaten