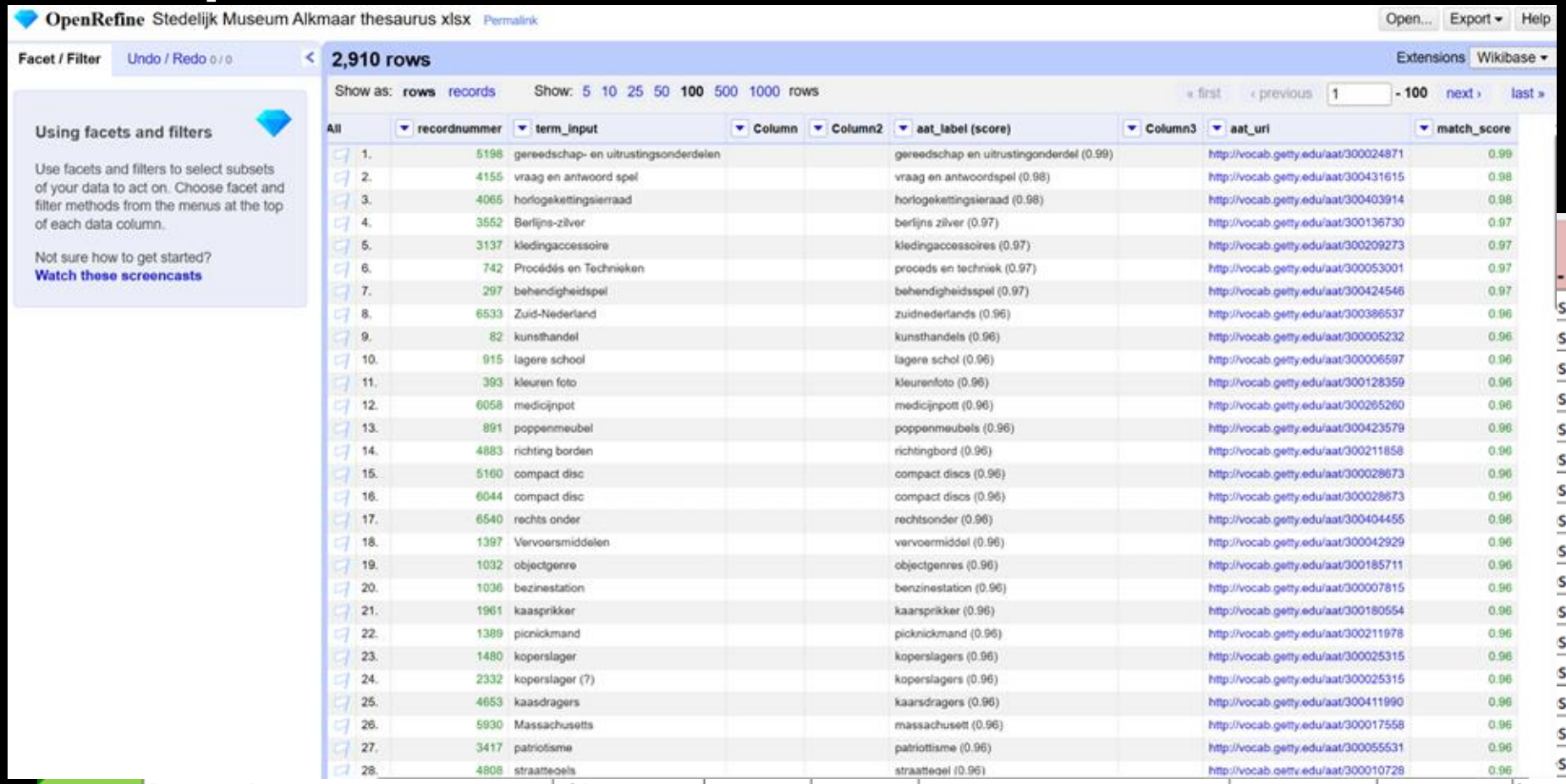
Nine Claassen
Netwerk Digitaal Erfgoed

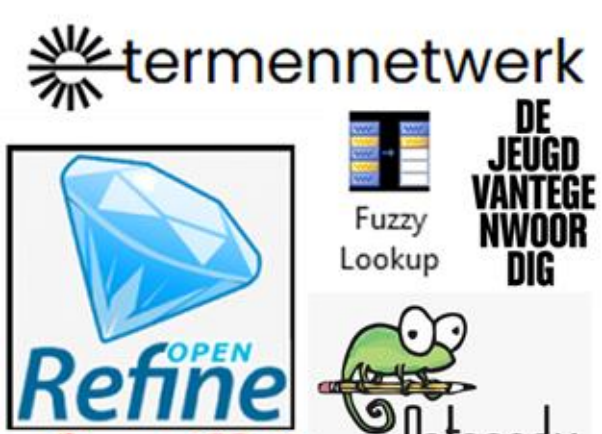
Shannon van Muijden
Rijksmuseum

2 datadudes
11 erfgoed datasets...

Uitdaging: zoveel mogelijk data schonen in één dag

Aanpak





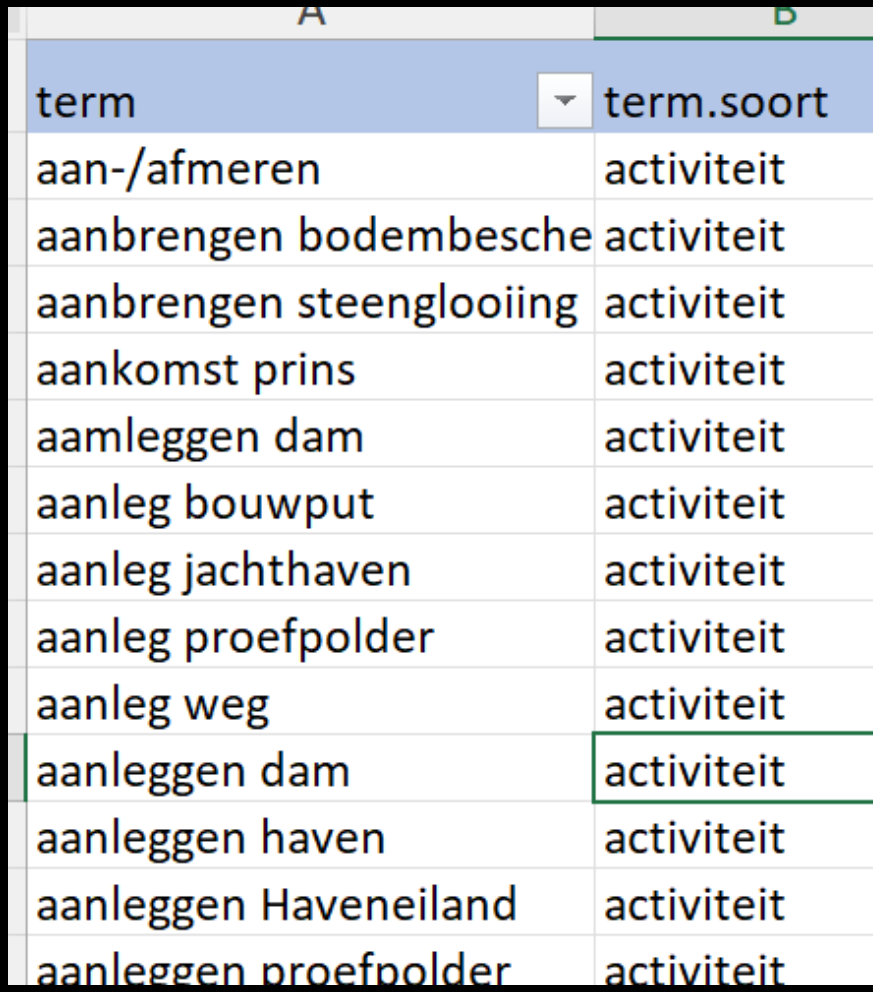
1393	Zeemeermin	objectnaam						https://data.cultureelerfgoed.nl/term/id/cht/27ac28e3-e69b-4bac
12	stoomfluit	objectnaam						https://data.cultureelerfgoed.nl/term/id/cht/292d741e-3a45-46ce
129	Informatieborden	objectcategorie		veiligheid|informatiebord				https://data.cultureelerfgoed.nl/term/id/cht/2e44391b-0677-4f76
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

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 steenglooiing	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

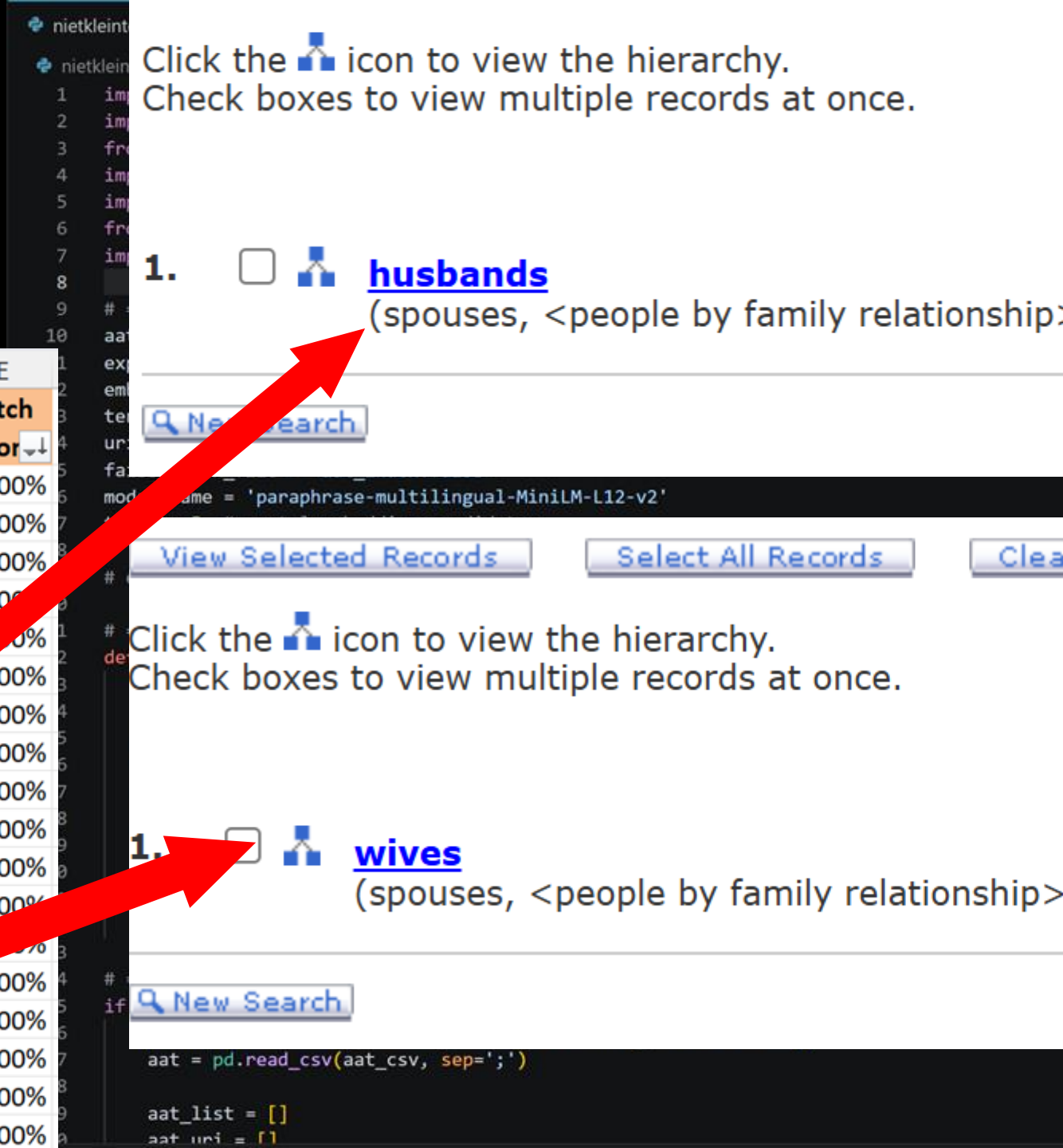
Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows « first ‹ previous 1 - 10 next › last »

All		recordnummer	term_input	Column	Column2	aat_label (score)	Column3	aat_uri	match_score
1.		5198	gereedschap- en uitrustingsonderdelen			gereedschap en uitrustingonderdel (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			proceds 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			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 h	aan-/afmeren	
3	aanbrengen bodembesche	activiteit			probeer de inhoudelijke informatie bij h	aanbrengen	
4	aanbrengen steenglooiing	activiteit			probeer de inhoudelijke informatie bij h	aanbrengen	
5	aankomst prins	activiteit			probeer de inhoudelijke informatie bij h	aankomen	
6	aamleggen dam	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
7	aanleg bouwput	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
8	aanleg jachthaven	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
9	aanleg proefpolder	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
10	aanleg weg	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
11	aanleggen dam	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
12	aanleggen haven	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
13	aanleggen Haveneiland	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
14	aanleggen proefpolder	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec
15	aanleggen strekdam	activiteit			probeer de inhoudelijke informatie bij h	aanleggen	https://data.cultureelerfgoed.nl/term/id/cht/51831506-aed7-43fe-a1f0-ecd9849f49ec

Vector embedding van de AAT

	A	B	C	D	E
1	record numr	term_input	aat_label (score)	aat_uri	match _scor
2	205	schenking	schenking (1.00)	aat:300138913	100%
3	3865	opdracht	opdracht (1.00)	aat:300026114	100%
4	371	jacht	jacht (1.00)	aat:300082988	100%
5	6424	varen	varen (1.00)	aat:300239493	10
6	377	veehoeder	veehoeder (1.00)	aat:300025615	0%
7	370	vissen	vissen (1.00)	aat:3001698	100%
8	7117	Bruinvis	bruinvis (1.00)	aat:3003 37	100%
9	7245	catalogusnummer	catalogusnummer (1.00)	aat:3 04620	100%
10	6574	inventarisnummer	inventarisnummer (1.00)	a 00312355	100%
11	148	fantasie	fantasie (1.00)	at:300068545	100%
12	147	hond	hond (1.00)	aat:300310627	100%
13	5229	constructie	constructie (1.00)	aat:300047195	100%
14	4466	vierkant	vierkant (1.00)	aat:300055637	
15	764	bedrijf	bedrijf (1.0	aat:3 4	100%
16	244	echtgenoot	echtgenoot (1.00)	00154341	100%
17	247	echtgenote	echtgenote (1.00)	aat:300154343	100%
18	233	eigenaar	eigenaar (1.00)	aat:300203630	100%
19	3071	gids	gids (1.00)	aat:300026300	100%



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