From Flat Lists to Taxonomies: Bottom-up Concept Scheme Generation in Linked Statistical Data

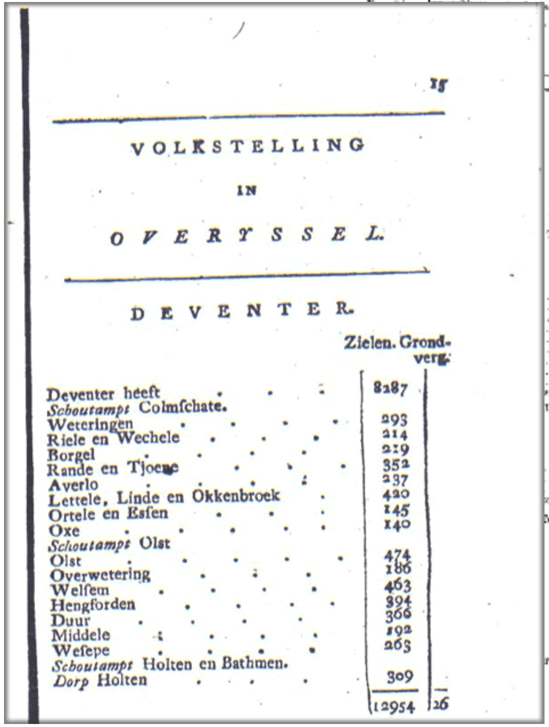
Albert Meroño-Peñuela
@albertmeronyo

SemStats @ ISWC 19-10-2014

Data Archiving and Networked Services
DANS
VU UNIVERSITY AMSTERDAM
eHumanities Royal Netherlands Academy of Arts and Sciences

Motivation

Dutch Historical Censuses (1795-1971)
[Public Historical Statistical Data]



15

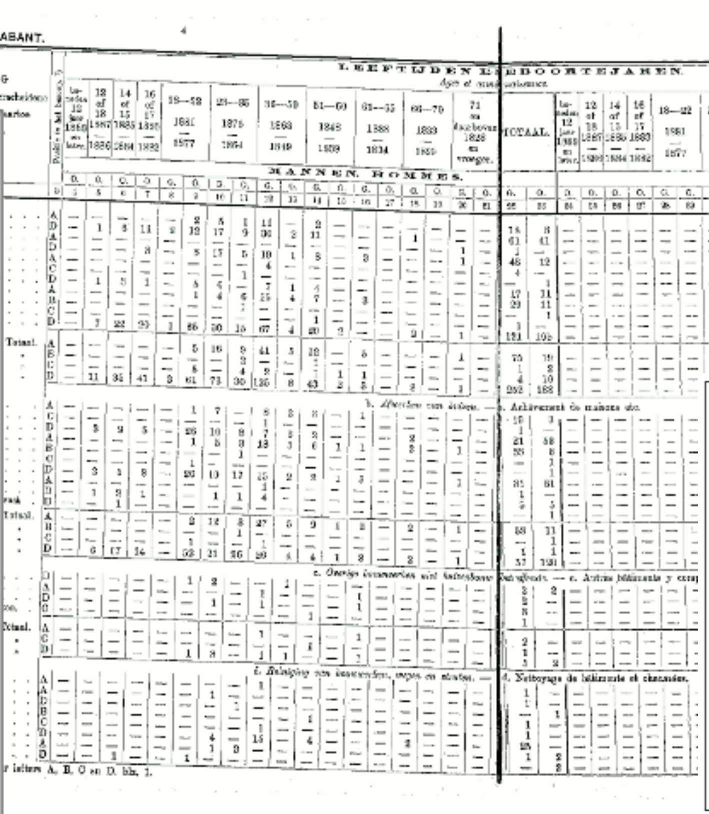
VOLKSTELLING

IN

OVERYSSEL.

DEVENTER.

	Zielen.	Grond-verg.
Deventer heeft	8287	
Schoutampt Colmſchate.		
Weteringen	293	
Riele en Wechele	214	
Borgel	219	
Rande en Tjoene	352	
Averlo	237	
Lettele, Linde en Okkenbroek	420	
Ortele en Esſen	145	
Oxe	140	
Schoutampt Olst		
Olst	474	
Overwetering	186	
Welſem	463	
Hengforden	394	
Duur	366	
Middele	192	
Weſepe	263	
Schoutampt Holten en Bathmen.		
Dorp Holten	309	
	12954	26



PROVINCIE NOORDBRABANT.

GEMEENTE BREDA.

BENAMING van de onderdeelen der onderscheiden

I. LEEFTIJDEN EN GEBOORTEJAREN.

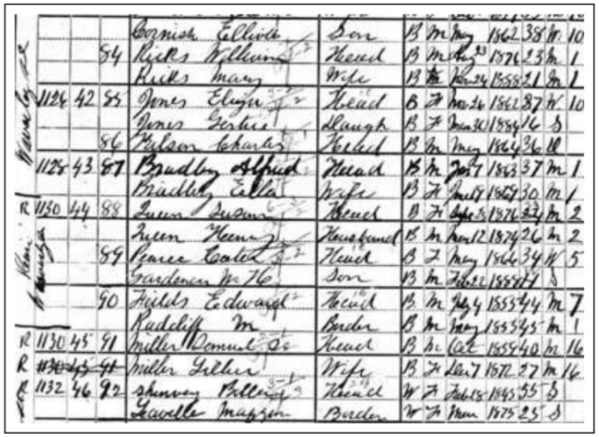
MANNEN. HOMMES.

VROUWEN. FEMMES.

TOTAAL.

TOTAAL van MANNEN en VROUWEN.

Totaal.



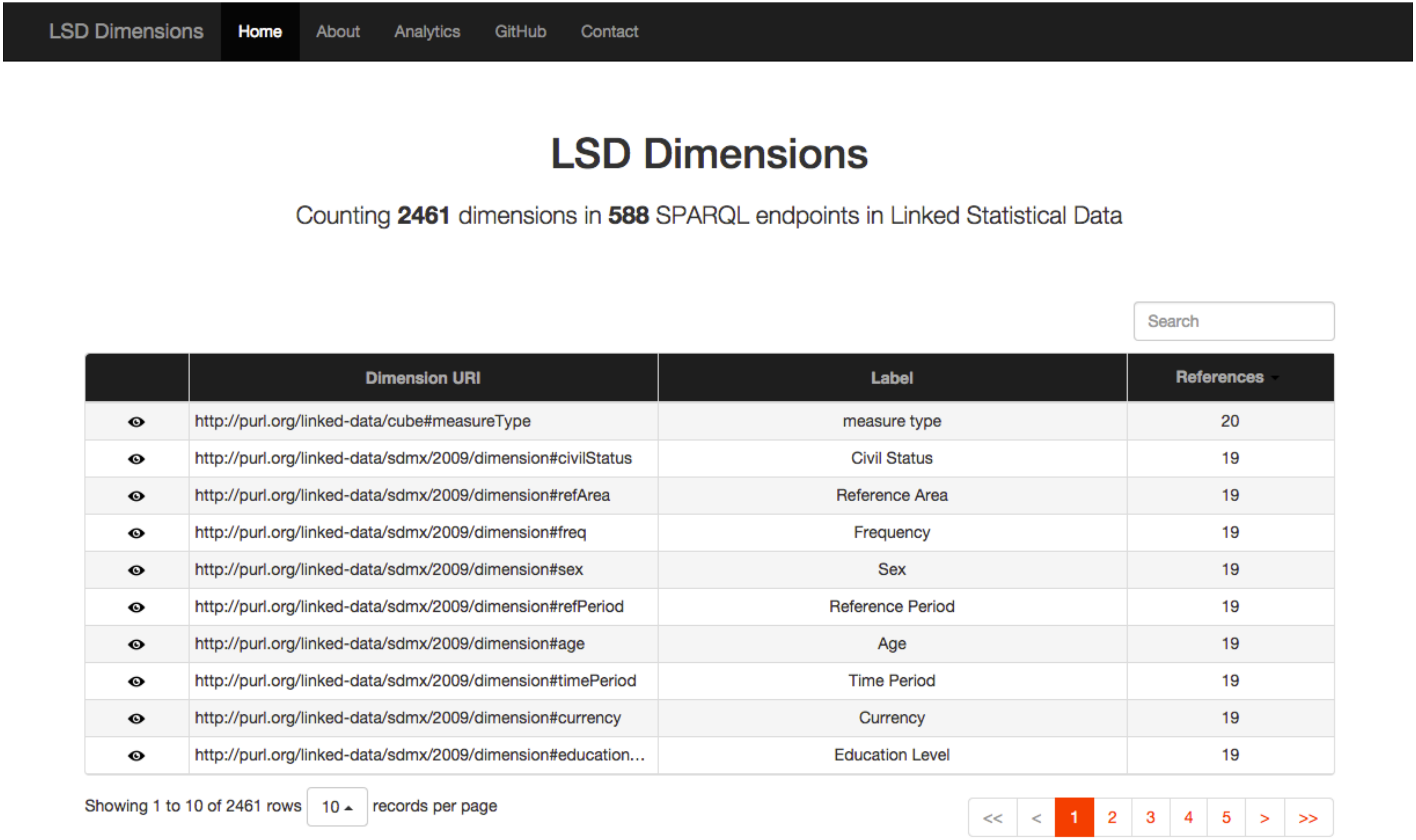
			Cornish Elliott	Son	B	M	May	1862	38	W	10
		84	Ricks William	Head	B	M	Aug 23	1876	23	M	1
			Ricks Mary	Wife	B	F	Jan 24	1858	21	M	1
1124	42	85	Jones Eliza	Head	B	F	Mar 26	1862	37	W	10
			Jones Gertie	Daugh	B	F	Mar 30	1884	16	S	
		86	Wilson Charles	Head	B	M	May	1864	36	W	
1125	43	87	Bradley Alfred	Head	B	M	Jan 7	1863	37	M	1
			Bradley Ella	Wife	B	F	Mar 17	1869	30	M	1
R 1130	44	88	Green Susan	Head	B	F	Apr 5	1876	24	M	2
			Green Henry	Husband	B	M	Nov 12	1874	26	M	2
		89	Pearce Caleb	Head	B	F	May	1866	34	W	5
			Gardener M T6	Son	B	M	Feb 22	1885	14	S	
		90	Fields Edward	Head	B	M	July 4	1855	44	M	7
			Radcliff M	Border	B	M	May	1855	45	M	1
R 1130	45	91	Miller Samuel Sr	Head	B	M	Oct	1859	40	M	16
R ~~1130~~	~~45~~	~~91~~	Miller Lillie	Wife	B	F	Dec 7	1872	27	M	16
R 1132	46	92	Shumway Billing	Head	W	F	Feb 15	1845	55	S	
			Lavelle Maggie	Border	W	F	Mar	1875	25	S	

http://lod.cedar-project.nl/cedar/

Dimension Reusability

```
cedar:BRT_1889_02_T1-S0-K17-h a qb:Observation ;
    cedar:population "12"^^xml:integer ;
    maritalstatus:maritalStatus
        maritalstatus:single ;
    cedarterms:occupationPosition cedarterms:job-D ;
    sdmx-dimension:sex sdmx-code:sex-F ;
    cedarterms:occupation hisco:88030 ;
    sdmx-dimension:refArea gg:11150 ;
    cedarterms:belief cedar:118 ;
    cedarterms:houseType cedar:Klooster ;
    prov:wasDerivedFrom
        cedar:BRT_1889_08_T1-S0-K17 ;
    prov:wasGeneratedBy
        cedar:BRT_1889_08_T1-S0-K17-activity .
```

LSD Dimensions



LSD Dimensions

Counting **2461** dimensions in **588** SPARQL endpoints in Linked Statistical Data

Search

	Dimension URI	Label	References
👁	http://purl.org/linked-data/cube#measureType	measure type	20
👁	http://purl.org/linked-data/sdmx/2009/dimension#civilStatus	Civil Status	19
👁	http://purl.org/linked-data/sdmx/2009/dimension#refArea	Reference Area	19
👁	http://purl.org/linked-data/sdmx/2009/dimension#freq	Frequency	19
👁	http://purl.org/linked-data/sdmx/2009/dimension#sex	Sex	19
👁	http://purl.org/linked-data/sdmx/2009/dimension#refPeriod	Reference Period	19
👁	http://purl.org/linked-data/sdmx/2009/dimension#age	Age	19
👁	http://purl.org/linked-data/sdmx/2009/dimension#timePeriod	Time Period	19
👁	http://purl.org/linked-data/sdmx/2009/dimension#currency	Currency	19
👁	http://purl.org/linked-data/sdmx/2009/dimension#education...	Education Level	19

Showing 1 to 10 of 2461 rows 10 ▴ records per page

http://lsd-dimensions.org/

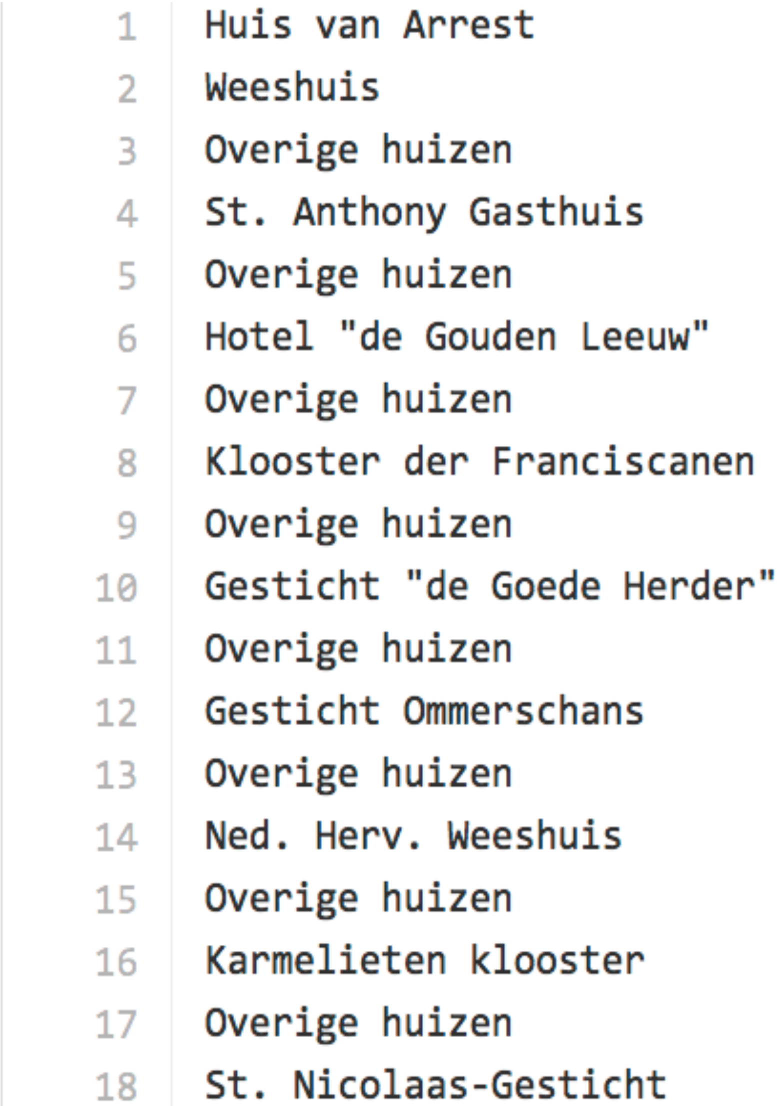
https://github.com/albertmeronyo/LSD-Dimensions

Hourly JSON-LD dumps

What if dimensions aren’t out there?

- Need to build them
- Input: flat lists of non-standard values
- Output: standard concept scheme
- Knowledge intensive problem

Dutch Historical Building Types



```
Huis van Arrest
Weeshuis
Overige huizen
St. Anthony Gasthuis
Overige huizen
Hotel "de Gouden Leeuw"
Overige huizen
Klooster der Franciscanen
Overige huizen
Gesticht "de Goede Herder"
Overige huizen
Gesticht Ommerschans
Overige huizen
Ned. Herv. Weeshuis
Overige huizen
Karmelieten klooster
Overige huizen
St. Nicolaas-Gesticht
```

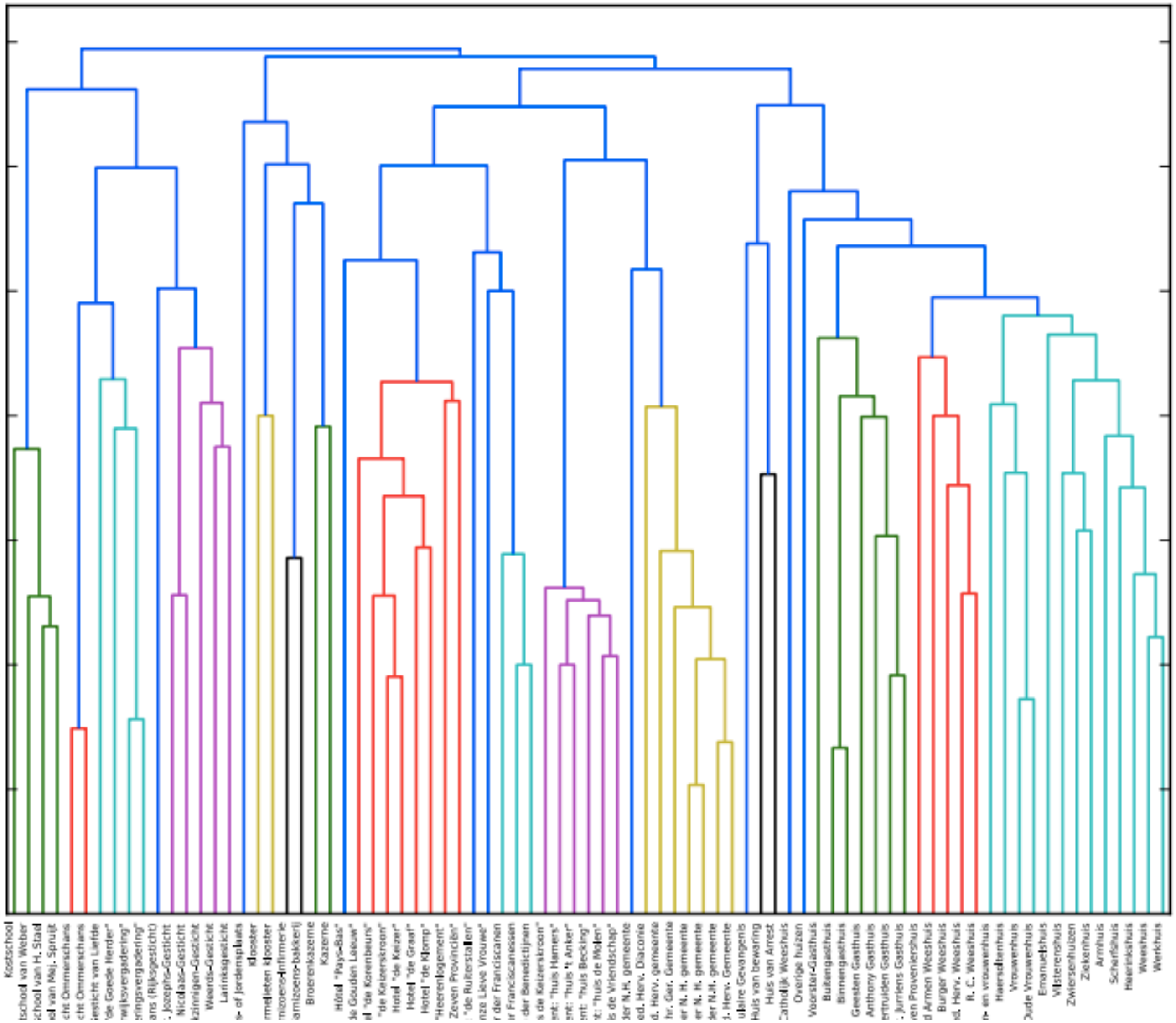
TabCluster: data-driven concept scheme generation

https://github.com/CEDAR-project/TabCluster

Leverages lexical and semantic term properties

- Lexical Hierarchical Clustering
  - ScyPy python implementation
  - Similarity transitivity
  - Clustering threshold & other parameters
- Semantic Tagging of Clusters (WordNet & DBPedia) w/ skos:Concept
  - Term frequency-based
  - Less common broader concept category

Output & Evaluation



Thank you

Questions, suggestions, comments most welcome

@albertmeronyo

https://github.com/CEDAR-project/TabCluster
http://lsd-dimensions.org/