

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



eHumanities
Royal Netherlands Academy of Arts and Sciences

Motivation

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

PROVINCIE NOORDBRABANT.												GENEENTE BREDA.											
B E N A M I N G van de volkstelling der inwonenden in de gemeente Breda.												B E N A M I N G van de volkstelling der inwonenden in de gemeente Breda.											
D. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.												D. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.											
A. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.												A. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.											
B. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.												B. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.											
C. 3. 3. 3. 3. 3. 3. 3. 3. 3. 3. 3. 3.												C. 3. 3. 3. 3. 3. 3. 3. 3. 3. 3. 3. 3.											
D. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4.												D. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4.											
E. 5. 5. 5. 5. 5. 5. 5. 5. 5. 5. 5. 5.												E. 5. 5. 5. 5. 5. 5. 5. 5. 5. 5. 5. 5.											
F. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6.												F. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6.											
G. 7. 7. 7. 7. 7. 7. 7. 7. 7. 7. 7. 7.												G. 7. 7. 7. 7. 7. 7. 7. 7. 7. 7. 7. 7.											
H. 8. 8. 8. 8. 8. 8. 8. 8. 8. 8. 8. 8.												H. 8. 8. 8. 8. 8. 8. 8. 8. 8. 8. 8. 8.											
I. 9. 9. 9. 9. 9. 9. 9. 9. 9. 9. 9. 9.												I. 9. 9. 9. 9. 9. 9. 9. 9. 9. 9. 9. 9.											
J. 10. 10. 10. 10. 10. 10. 10. 10. 10. 10. 10. 10.												J. 10. 10. 10. 10. 10. 10. 10. 10. 10. 10. 10. 10.											
K. 11. 11. 11. 11. 11. 11. 11. 11. 11. 11. 11. 11.												K. 11. 11. 11. 11. 11. 11. 11. 11. 11. 11. 11. 11.											
L. 12. 12. 12. 12. 12. 12. 12. 12. 12. 12. 12. 12.												L. 12. 12. 12. 12. 12. 12. 12. 12. 12. 12. 12. 12.											
M. 13. 13. 13. 13. 13. 13. 13. 13. 13. 13. 13. 13.												M. 13. 13. 13. 13. 13. 13. 13. 13. 13. 13. 13. 13.											
N. 14. 14. 14. 14. 14. 14. 14. 14. 14. 14. 14. 14.												N. 14. 14. 14. 14. 14. 14. 14. 14. 14. 14. 14. 14.											
O. 15. 15. 15. 15. 15. 15. 15. 15. 15. 15. 15. 15.												O. 15. 15. 15. 15. 15. 15. 15. 15. 15. 15. 15. 15.											
P. 16. 16. 16. 16. 16. 16. 16. 16. 16. 16. 16. 16.												P. 16. 16. 16. 16. 16. 16. 16. 16. 16. 16. 16. 16.											
Q. 17. 17. 17. 17. 17. 17. 17. 17. 17. 17. 17. 17.												Q. 17. 17. 17. 17. 17. 17. 17. 17. 17. 17. 17. 17.											
R. 18. 18. 18. 18. 18. 18. 18. 18. 18. 18. 18. 18.												R. 18. 18. 18. 18. 18. 18. 18. 18. 18. 18. 18. 18.											
S. 19. 19. 19. 19. 19. 19. 19. 19. 19. 19. 19. 19.												S. 19. 19. 19. 19. 19. 19. 19. 19. 19. 19. 19. 19.											
T. 20. 20. 20. 20. 20. 20. 20. 20. 20. 20. 20. 20.												T. 20. 20. 20. 20. 20. 20. 20. 20. 20. 20. 20. 20.											
U. 21. 21. 21. 21. 21. 21. 21. 21. 21. 21. 21. 21.												U. 21. 21. 21. 21. 21. 21. 21. 21. 21. 21. 21. 21.											
V. 22. 22. 22. 22. 22. 22. 22. 22. 22. 22. 22. 22.												V. 22. 22. 22. 22. 22. 22. 22. 22. 22. 22. 22. 22.											
W. 23. 23. 23. 23. 23. 23. 23. 23. 23. 23. 23. 23.												W. 23. 23. 23. 23. 23. 23. 23. 23. 23. 23. 23. 23.											
X. 24. 24. 24. 24. 24. 24. 24. 24. 24. 24. 24. 24.												X. 24. 24. 24. 24. 24. 24. 24. 24. 24. 24. 24. 24.											
Y. 25. 25. 25. 25. 25. 25. 25. 25. 25. 25. 25. 25.												Y. 25. 25. 25. 25. 25. 25. 25. 25. 25. 25. 25. 25.											
Z. 26. 26. 26. 26. 26. 26. 26. 26. 26. 26. 26. 26.												Z. 26. 26. 26. 26. 26. 26. 26. 26. 26. 26. 26. 26.											
Aanmerkingen van de volkstelling.												Aanmerkingen van de volkstelling.											
Deventer heeft Schoutenstamps Colmschate.												Cognac, Ellwood, Son											
Weteringen												54 Ryckels, William, Head											
Riele en Wechele												Hicks, Mary, Wife											
Borgel												1124 42 88 Jones, Eliza, Head											
Rande en Tjoeeme												Jones, Gertrude, Daughter											
Averlo												56 Palser, Charles, Head											
Lettele, Linde en Okkenbroek												1128 43 87 Bradley, Alfred, Head											
Ortele en Esfen												Bradley, Alice, Wife											
Oxe												R 1130 44 88 Juan, Susanna, Head											
Schoutenstamps Oist												Juan, Henry, Husband											
Oist												89 Juan, Cecilia, Head											
Overwetering												Gardiner, John, Son											
Welfem												90 Ryckels, Edward, Head											
Hengforde												Ryckels, M., Brother											
Duur												R 1130 45 91 Miller, Manuel, Head											
Middelse												R 1130 45 91 Miller, Samuel, Head											
Wepepe												R 1132 46 92 Murphy, Phillip, Head											
Schoutenstamps Holten en Bathmen.												Murphy, Maggie, Wife											
Dorp Holten												Murphy, Tom, Son											
827												12954 26											
D E V E N T E R.												c. Aanmerkingen van huishoudens.											
Zielien, Grond- verg.												d. Aanmerkingen van huishoudens.											
Deventer heeft												e. Aanmerkingen van huishoudens en huishoudendeelgenoten.											
Schoutenstamps Colmschate.												f. Aanmerkingen van huishoudendeelgenoten.											
Weteringen												g. Nettogetal van huishoudendeelgenoten.											
Riele en Wechele												h. Nettogetal van huishoudendeelgenoten.											
Borgel												i. Nettogetal van huishoudendeelgenoten.											
Rande en Tjoeeme												j. Nettogetal van huishoudendeelgenoten.											
Averlo												k. Nettogetal van huishoudendeelgenoten.											
Lettele, Linde en Okkenbroek												l. Nettogetal van huishoudendeelgenoten.											
Ortele en Esfen												m. Nettogetal van huishoudendeelgenoten.											
Oxe												n. Nettogetal van huishoudendeelgenoten.											
Schoutenstamps Oist												o. Nettogetal van huishoudendeelgenoten.											
Oist												p. Nettogetal van huishoudendeelgenoten.											
Overwetering												q. Nettogetal van huishoudendeelgenoten.											
Welfem												r. Nettogetal van huishoudendeelgenoten.											
Hengforde												s. Nettogetal van huishoudendeelgenoten.											
Duur												t. Nettogetal van huishoudendeelgenoten.											
Middelse												u. Nettogetal van huishoudendeelgenoten.											
Wepepe												v. Nettogetal van huishoudendeelgenoten.											
Schoutenstamps Holten en Bathmen.												w. Nettogetal van huishoudendeelgenoten.											
Dorp Holten												x. Nettogetal van huishoudendeelgenoten.											
309												y. Nettogetal van huishoudendeelgenoten.											
12954												z. Nettogetal van huishoudendeelgenoten.											

<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

Home

About

Analytics

GitHub

Contact

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#educationLevel	Education Level	19

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

<< < 1 2 3 4 5 > >>

<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

- | | |
|----|----------------------------|
| 1 | Huis van Arrest |
| 2 | Weeshuis |
| 3 | Overige huizen |
| 4 | St. Anthony Gasthuis |
| 5 | Overige huizen |
| 6 | Hotel "de Gouden Leeuw" |
| 7 | Overige huizen |
| 8 | Klooster der Franciscanen |
| 9 | Overige huizen |
| 10 | Gesticht "de Goede Herder" |
| 11 | Overige huizen |
| 12 | Gesticht Ommerschans |
| 13 | Overige huizen |
| 14 | Ned. Herv. Weeshuis |
| 15 | Overige huizen |
| 16 | Karmelieten klooster |
| 17 | Overige huizen |
| 18 | St. Nicolaas-Gesticht |

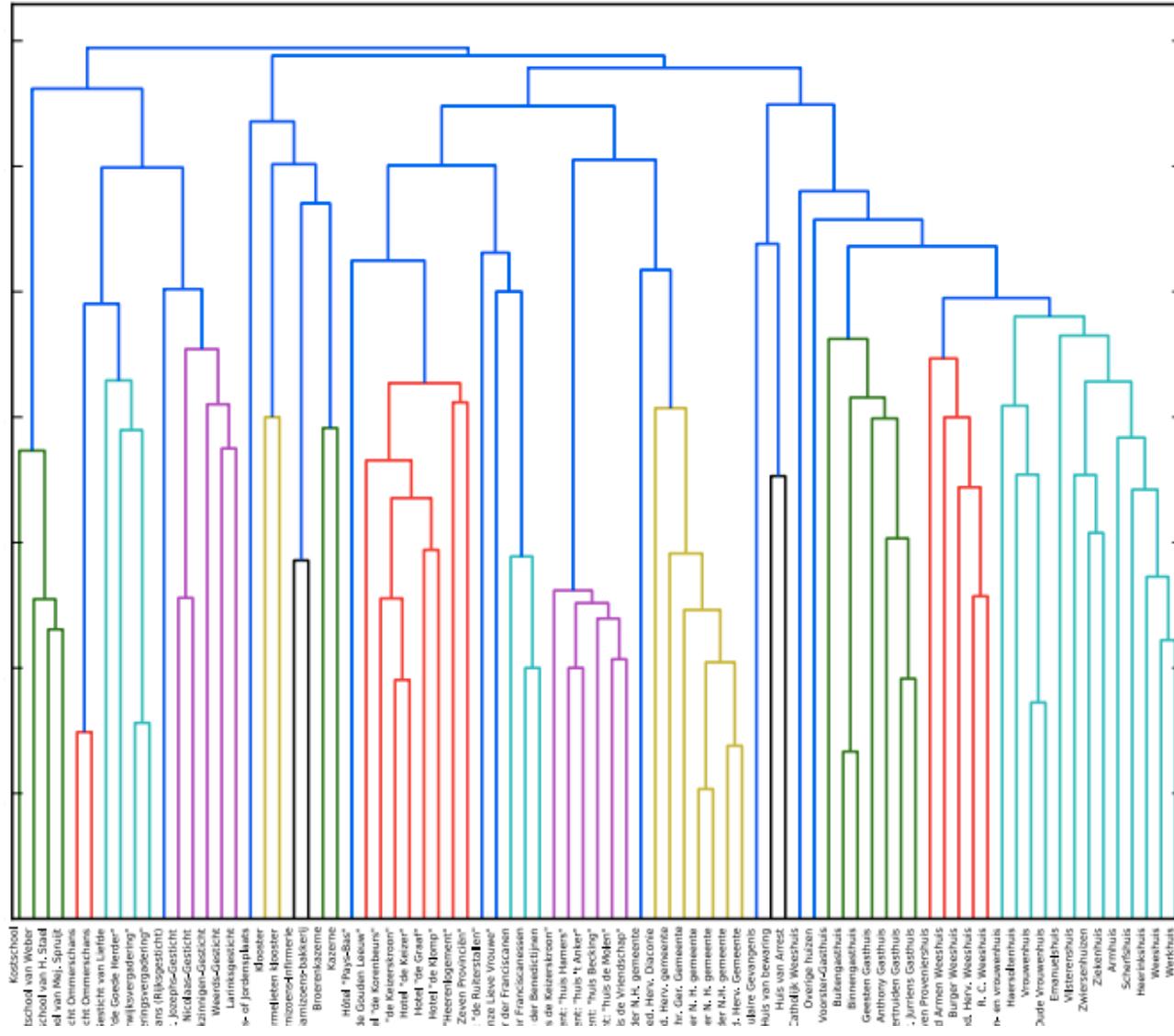
TabCluster: data-driven concept scheme generation

<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/>