Detecting and Reporting Extensional Concept Drift in Statistical Linked Data

Albert Meroño-Peñuela[1,2] Christophe Guéret[2]
Rinke Hoekstra[1,3] Stefan Schlobach[1]

[1]Department of Computer Science, VU University Amsterdam, NL

[2]Data Archiving and Networked Services, KNAW, NL

[3]Leibniz Center for Law, Faculty of Law, University of Amsterdam, NL

October 22th, 2013
First International Workshop on Semantic Statistics
ISWC 2013

As the world changes continuously, concepts also change their meaning over time

We call this *concept drift*

- Smooth transitions
- Radical transitions (*concept shift*)

Concepts are also present in SLD

- Variable meaning/semantics (what the variable is supposed to represent?)
- Variable values/factors (*RomschKatholik*, *RomsKatholic*, *KatholicChristelijk*)

To what extent stability of meaning of these concepts is guaranteed in data collected on *very long time ranges*?

If meaning is stable

- Old models are reusable
- Backwards comparisons always make sense

Else

- New models may be necessary
- Backwards comparisons may be incorrect

The meaning of a concept C can change in several ways[1].

- **Intension drift** occurs when there is a difference in the *properties or attributes* of two variants of the same concept ($sim_{int}(C', C'') \neq 1$).
- **Extension drift** occurs when there is a difference in the *individuals that belong* to two variants of the same concept ($sim_{ext}(C', C'') \neq 1$).
- **Label drift** occurs when there is a difference in the *labels* of two variants of the same concept ($sim_{label}(C', C'') \neq 1$).

[1] S. Wang, S. Schlobach, M. Klein, *What Is Concept Drift and How to Measure It?*, 2010

Key ideas
If you remember nothing else, remember this

In this talk we present two key ideas

- A method to detect extensionally drifted concepts
- A recommendation to annotate such drifts in SLD

- We define the *extension function* $ext(C)$ of a concept C as the **number of individuals** that belong to C

$$ext(C) = |\{a : C\}|$$

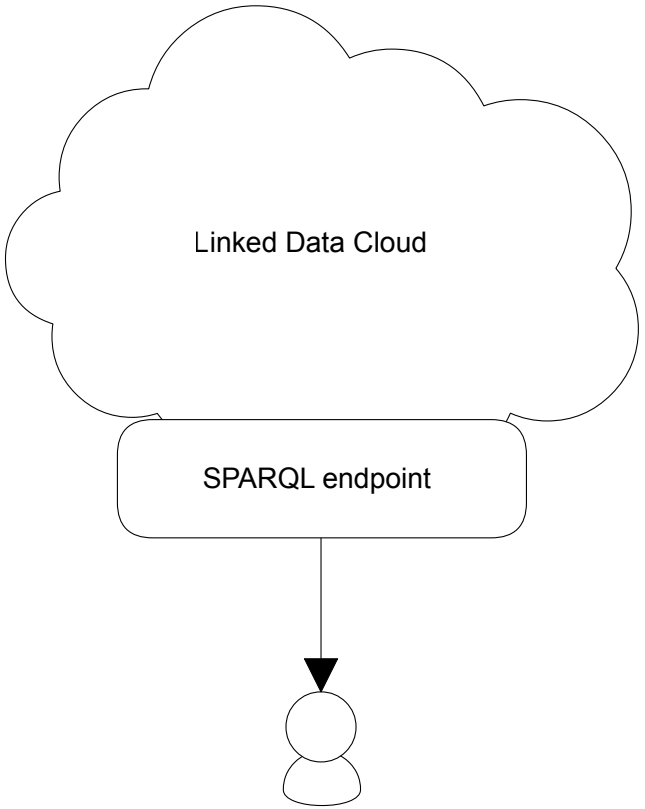
- We define the *extension similarity function* $sim_{ext}(C', C'')$ between two variants C', C'' of a concept C as the function that returns the **probability that C' and C'' have identical populations**[2]

$$sim_{ext}(C', C'') = wilcox.test(ext(C'), ext(C''))$$

- Intuitive idea: concept variants with significantly different populations ($p < 0.05$) suffer radical transitions

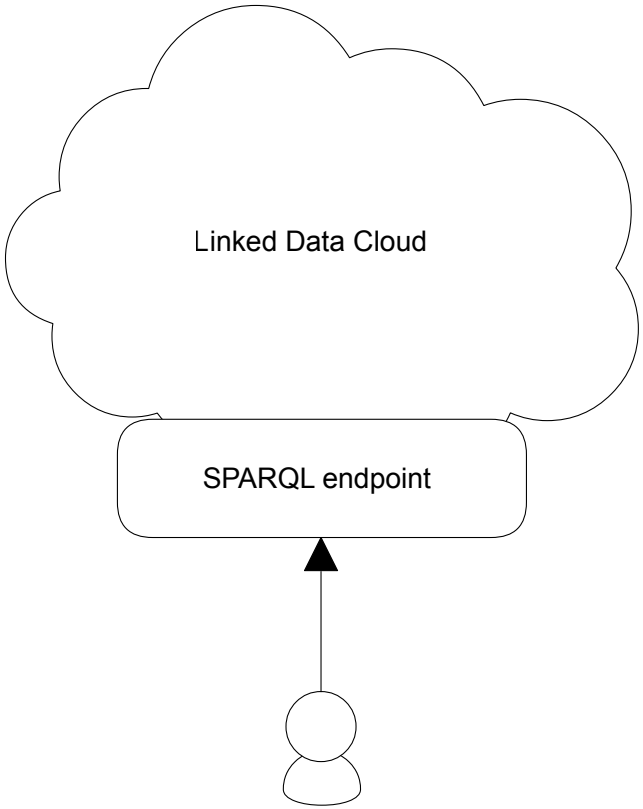
[2] F. Wilcoxon, *Individual comparisons by ranking methods*, 1945

Extensional drift reporting
Using SPARQL UPDATEs



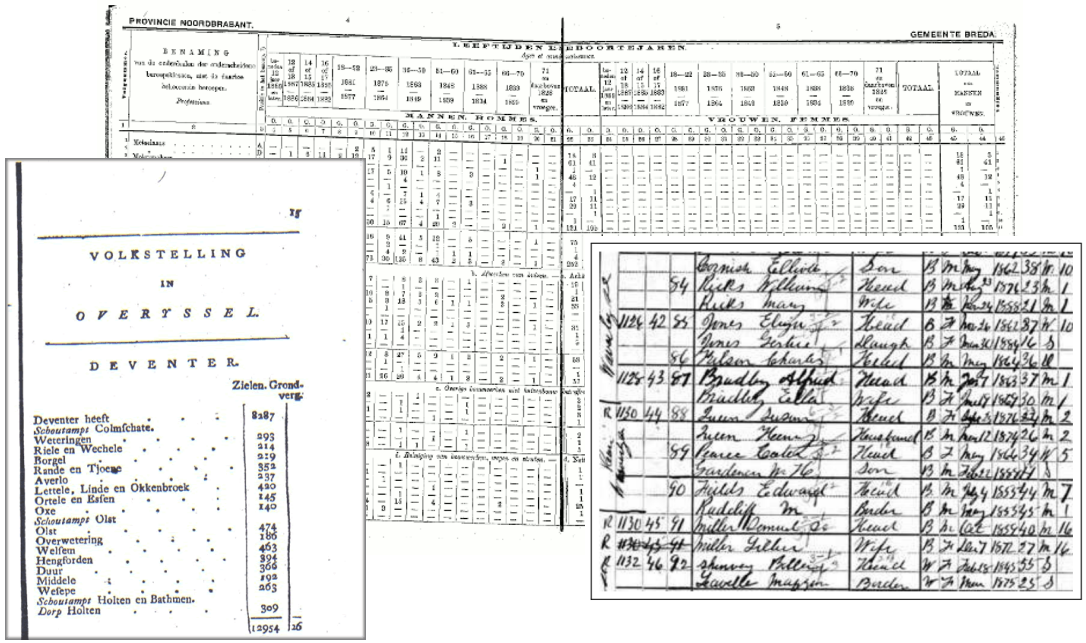
Typical analysis workflow. Users SELECT data from SLD, but analyses run offline and results are not pushed back

Extensional drift reporting
Using SPARQL UPDATEs



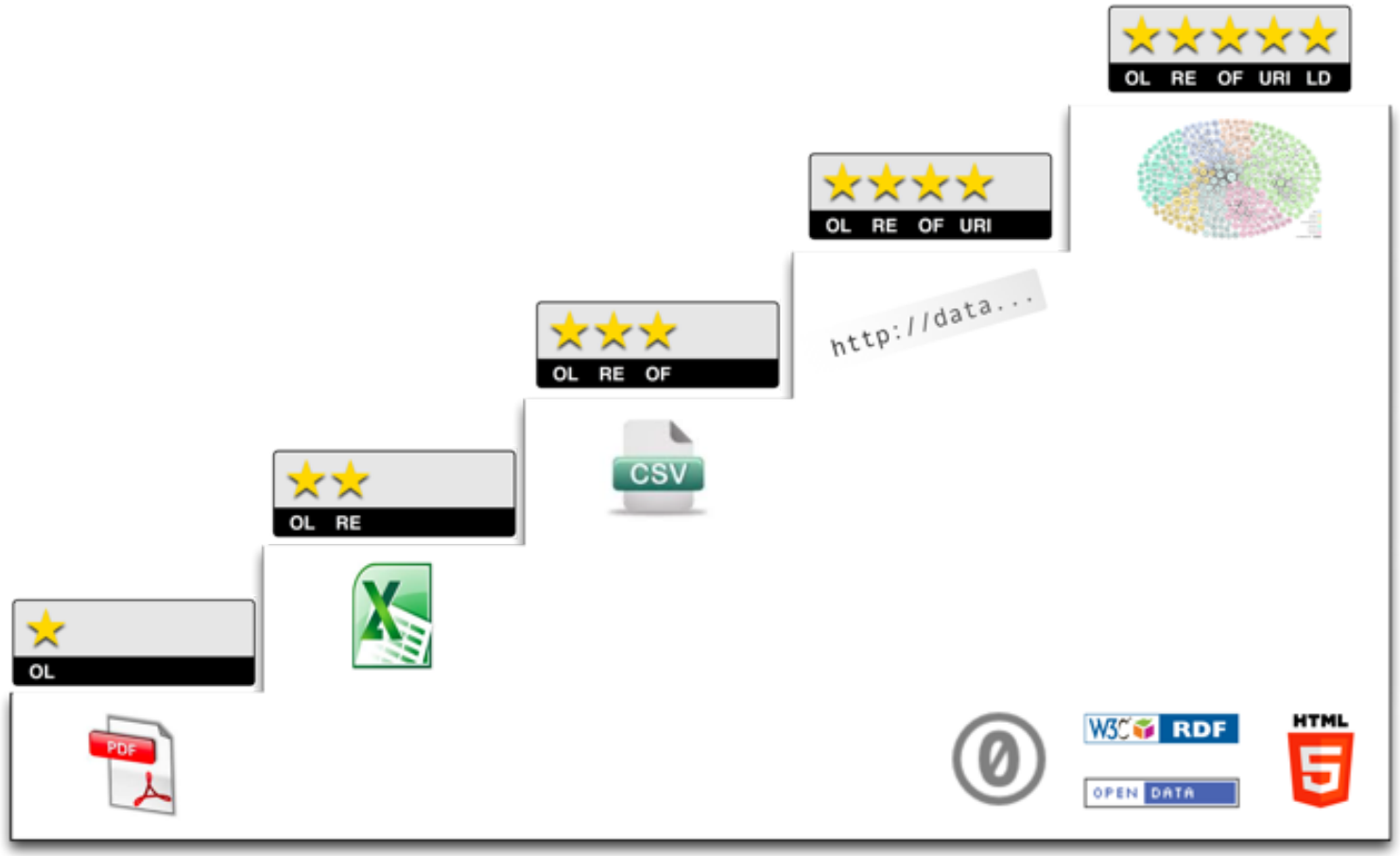
Proposal. After running analyses, UPDATE results (e.g. detected extensional drifts) back to the endpoint

The Dutch historical censuses (1795 - 1971)

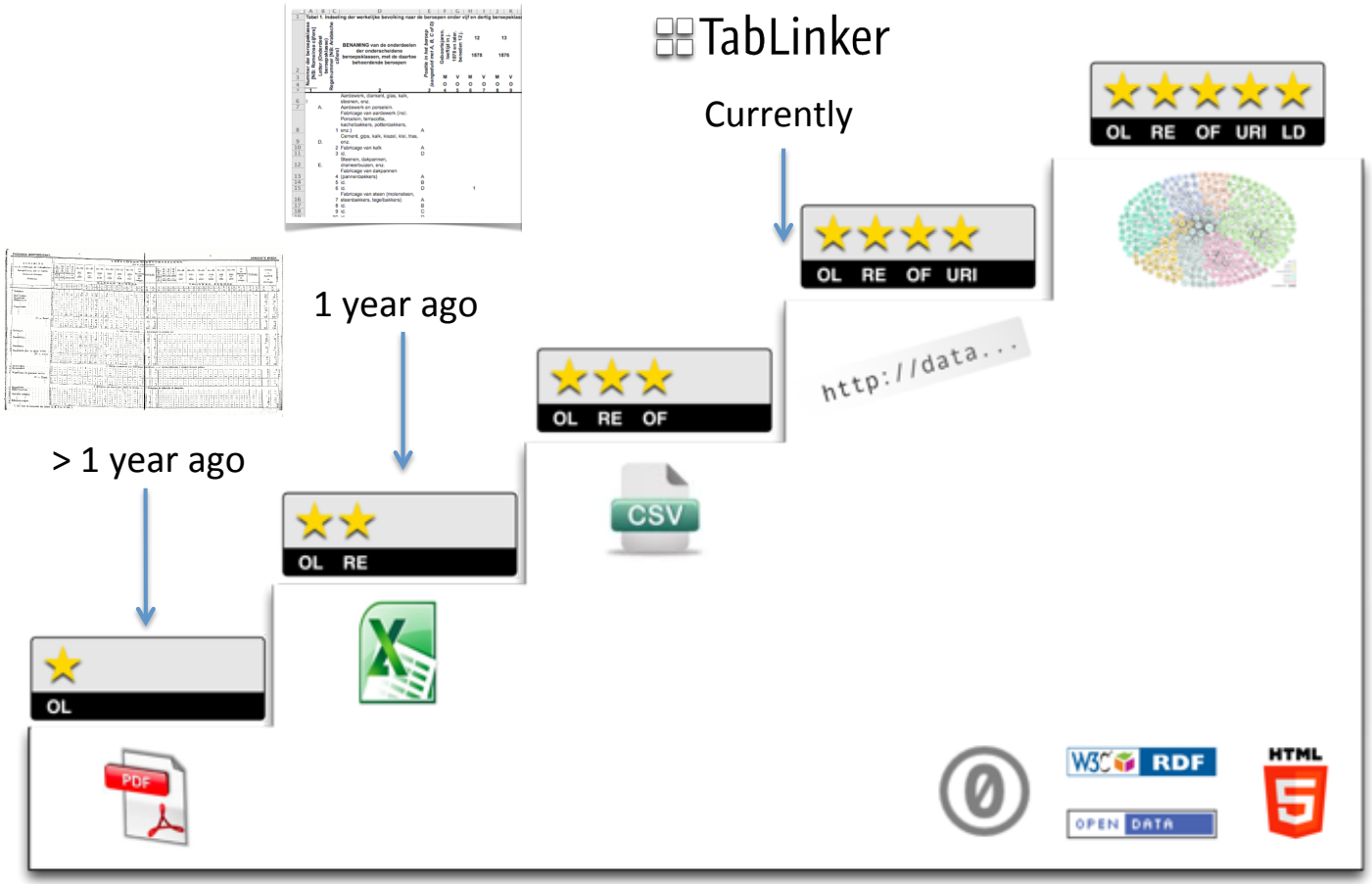


- Lots of statistical data (2,288 tables)
- Lots of concepts
- Big time span (176 years)

May contain lots of drifted concepts!



Towards 5-star historical census data



TabLinker[3] : Supervised XLS2RDF converter

[3] https://github.com/Data2Semantics/TabLinker

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX d2s: <http://www.data2semantics.org/core/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX ns: <row_property_URI>

SELECT ?d1label ... ?dnlabel ?p1label ... ?pmlabel ?population
FROM <named_graph_URI>
WHERE {
 ?cell d2s:isObservation [ a qb:Observation ;
                           qb:DimensionProperty ?d1 ... ?dn ;
                           ns:property1 ?p1 ;
                           ...
                           ns:propertym ?pm ;
                           qb:MeasureProperty ?population ] .
 OPTIONAL {
 ?cell d2s:isObservation [ns:propertyk ?pk ] .
 ?pk skos:prefLabel ?pklabel . }
 ...
 OPTIONAL {
 ?cell d2s:isObservation [qb:DimensionProperty ?dl ] .
 ?dl skos:prefLabel ?dllabel . }
 ?pt skos:broader ?pu .
 ?pu skos:broader ?pv .
 d1 ... dn skos:prefLabel ?d1label ... ?dnlabel .
 p1 ... pm skos:prefLabel ?p1label ... ?pmlabel .
 FILTER (?d1 IN (v1, ..., vr)) ...
 FILTER (?dn IN (w1, ..., ws))
}
```

SPARQL template for unfolding RDF Data Cubes

Census contains very heterogeneous data
Subset selection: occupation census of 1889 and 1899

Age range	Gender	Marital status	Municipality
36-50	M	G	Velsen
51-60	V	O	Zaandam
23-35	VROUWEN	O.	Haarlem
36-50	MANNEN	G.	Weesp

Class	Subclass	Occupation	Position	Population
II	b	Aanemers	A	3
IV	a	Agenten	B	1
XXI	d	Ambtenaren en beambten	C	2
I	h	Afwerken van huizen	A	5

1889 HISCO mappings

Arbeiders	D	99900
Kooplieden	A	41025
Winkeliers	A	41030
Handel in voorwerpen van kleedig.		41025
Kooplieden	A	41025
Winkeliers	A	41030
Winkelbedienden	C	45130
Handel in voorwerpen van voeding en genot.		41025
Arbeiders	C	99900
Arbeiders	D	99900
Dêpothouders	B	45130
Kooplieden	A	41025
Winkeliers	A	41030
Winkelbedienden	C	45130
Handel in voorwerpen van woning.		44130
Kooplieden	A	41025
Winkeliers	A	41030
Handel in boek- en kusntwerken (incl. dagbladen)		41030
Uitegevers	A	21110
Winkeliers	A	41030
Handel in luxe artikelen.		41025
Kooplieden	B	41025
Winkeliers	A	41030
Handel in levend vee en gevogelte		41025
Kooplieden	A	41025
Handel in andere waren.		41025
Arbeiders	D	99900
Kooplieden	A	41025
Kramers en rondventers	A	45220
Kramers en rondventers	D	45220
Magazijn- en pakhuisknechts	D	97145
Winkeliers	A	41030

1899 HISCO mappings

Schaapherders	D	62430
Hoenderfokkers	A	61260
Hoenderfokkers	D	61260
Vogelkweekers	A	61290
Bijenhouders	A	61290
Bijenhouders	D	61290
Boterboeren	A	77530
Boterboeren	B	77530
Boterboeren	D	77530
Melkboeren (niet melkslijters)	B	41030
Beestensnijders	A	77330
Hoenderparkhouders	A	61260
Tuinlieden	B	62740
Bloembollenkweekers	B	61270
Boomkweekers	B	61230
Boomsnoeiers	C	62730
Nettenboeters	D	75465
Scheepslossers	A	97120
Scheepslossers	C	97120
Scheepslossers	D	97120
Schelpenvisschers	A	64990
Schelpenvisschers	B	64990
Schelpenvisschers	C	64990
Schelpenvisschers	D	64990
Wiervisschers	A	64990
Vischsnijders	D	77940
Visschers	B	64100
Visschers	C	64100
Jagers	D	64960
Eendenkooihouders	A	64960
Eendenkooihouders	D	64960
Pers. in algem. dienst	C	30000

HISCO normalization

Is there extensional concept drift between two variants of the same occupation (i.e. same HISCO code)?

E.g. is there concept drift between *ship loaders* of 1889 and 1899?

Are the occupations in the two censuses comparable?

- 217 common HISCO codes
- 72.2% of all 1889 HISCO codes
- 71.1% of all 1899 HISCO codes

For all h common HISCO codes

- Query for population distributions of h in 1889
- Query for population distributions of h in 1899
- Do `wilcox.test` between the two and get *p-values*

Case-Study
Results

HISCO	Occupation	p-value
97125	Loader of ship, truck, wagon or airplane	1.83e-10
21110	General manager	4.23e-09
41025	Working proprietor (wholesale, retail trade)	1.52e-08
79100	Tailor	7.75e-07
57030	Barber, hairdresser	1.17e-04
88010	Jeweller	1.84e-04

(a) Occupations with stronger ext. drift.

HISCO	Occupation	p-value
53190	Other cooks	1.00
75452	Lace weaver	1.00
75490	Other weavers	1.00
75990	Other spinners, weavers, knitters, dyers	1.00
77690	Other bakers, pastry cooks and confectionery makers	1.00

(b) Occupations with greater ext. stability.

Group	Type	p-value
7, 8, 9	Production, transport, operators	2.03e-19
5	Service workers	1.88e-12
4	Sales workers	2.94e-08
2	Administrative and managerial	4.20e-08

(c) Major groups with stronger ext. drift.

Group	Type	p-value
6	Agriculture, animal husbandry, fishermen, hunters	0.38
0, 1	Professional and technical	0.16
3	Clerical	1.40e-04

(d) Major groups with greater ext. stability.

Table 2: Wilcoxon test p-values per HISCO code ((a),(b)) and major group ((c),(d)).

- Late industralization of the Netherlands (late 19th century)
- 19.35% of occupations show radical extensional drifts

```
1  PREFIX d2s: <http://www.data2semantics.org/core/>
2  PREFIX d2s1889: <urn:nbn:nl:ui:13-m4k-4lp>
3  PREFIX d2s1899: <urn:nbn:nl:ui:13-988-0dq>
4
5  INSERT DATA {
6    GRAPH <named_graph_URI> {
7      d2s1889:Sjouwerlieden d2s:isDrift [
8        d2s:extDrift d2s1899:Expeditie_bevrachters_bestellers_sjouwerlieden ,
9                     d2s1899:Personeel_voor_laden_en_lossen ,
10                    d2s1899:Personeel_voor_lading_en_lossing ,
11                    d2s1899:Sjouwerlieden ;
12       d2s:weight 1.83e-10 ] . } }
```

Listing 1.3: Excerpt of the SPARQL query reporting back extensionally drifted occupational concepts. Only the drift for one occupational concept is shown. Inverse drifts from the second graph to the first are also issued.

Closing the pull-push cycle[4]

[4] https://github.com/albertmeronyo/ConceptDrift/

Conclusions & Future work

- Concept drift: concepts change over time
- It affects model reusability and backwards querying
- Extensional drifts can be detected with `wilcox.test`
- We SPARQL UPDATE drifts to let others know

- Scale up variables, tables
- Parametrization of the `wilcox.test` depending on the time gap
- Integration of intension and label drift measures
- RDF HISCO will be released soon
- HISCO mappings will be published as separate named graphs (so that you can link yours)

Thank you
Questions, suggestions?

@albertmeronyo
http://www.cedar-project.nl
http://www.data2semantics.org