

Linked Edit Rules

A Web Friendly Way of Checking Quality
of RDF Data Cubes

Albert Meroño-Peñuela, Christophe Guéret,
Stefan Schlobach

Data Archiving and Networked Services



UNIVERSITY
AMSTERDAM



eHumanities
Royal Netherlands Academy of Arts and Sciences

Motivation

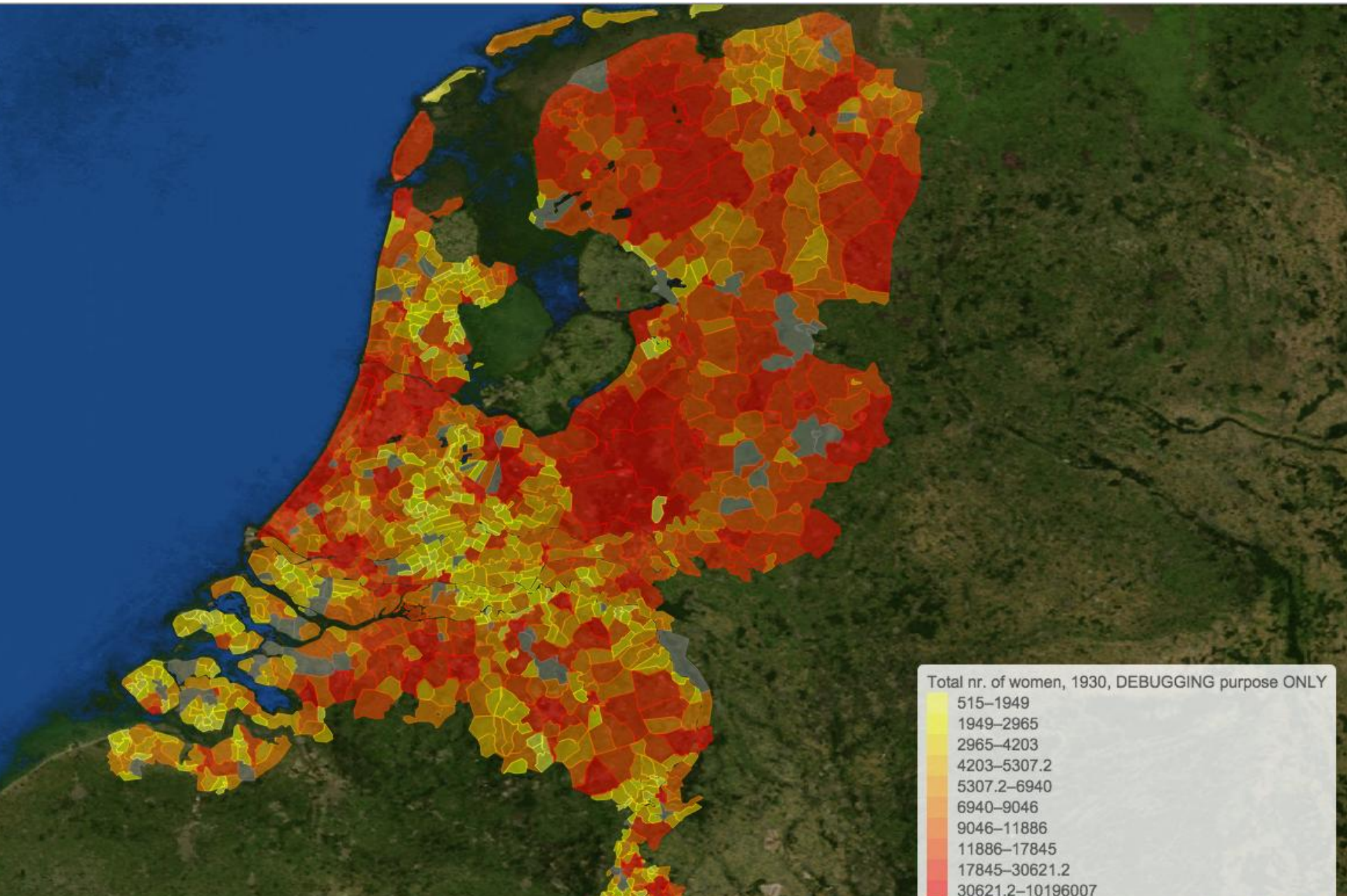
Dutch Historical Censuses (1795-1971) [Public Historical Statistical Data]

PROVINCIE NOORDBRABANT.		GENEENTE BREDA																	
BENAMING		MEEPTIJDEN EN BOORTEJAAREN														TOTAAL		TOTAAL	
van de oorspronkelijke oorspronkelijke		Ages at birth														Total		Total	
		MANNEN, HOMMES.							VROUWEN, FEMMES.										
		17-20	21-25	26-30	31-35	36-40	41-45	46-50	17-20	21-25	26-30	31-35	36-40	41-45	46-50	0-16	17-64	0-16	17-64
VOLKSTELLING IN OVERTYSSEL DEVENTER.		Zielen. Grondverg.																	
Deventer heeft		8287																	
Schoutstampt Colmfchate.		293																	
Weteringen		214																	
Riele en Wechele		219																	
Borgel		352																	
Rande en Tjoepe		237																	
Averlo		420																	
Lettele, Linde en Okkenbroek		145																	
Ortele en Esfen		140																	
Oxe		474																	
Schoutstampt Olst		188																	
Olst		463																	
Overwetering		394																	
Welfem		366																	
Hengforden		102																	
Duur		263																	
Middele		309																	
Wefeye		12954																	
Schoutstampt Holten en Bathmen.		126																	
Dorp Holten																			

A. Schiedamschen wijk.		B. Achterwijk de mannes wijk.	
C. Oude wijk van de mannes wijk.		D. Oude wijk van de mannes wijk.	
124	42	55	87
125	43	57	87
126	44	58	88
130	45	59	89
132	46	62	92

Handwritten	Handwritten	Handwritten	Handwritten	Handwritten
54	Reids	Wife	B. M. 1872	13. 1
55	Reids	Wife	B. M. 1874	13. 1
56	Reids	Wife	B. M. 1876	13. 1
57	Reids	Wife	B. M. 1878	13. 1
58	Reids	Wife	B. M. 1880	13. 1
59	Reids	Wife	B. M. 1882	13. 1
60	Reids	Wife	B. M. 1884	13. 1
61	Reids	Wife	B. M. 1886	13. 1
62	Reids	Wife	B. M. 1888	13. 1
63	Reids	Wife	B. M. 1890	13. 1
64	Reids	Wife	B. M. 1892	13. 1
65	Reids	Wife	B. M. 1894	13. 1
66	Reids	Wife	B. M. 1896	13. 1
67	Reids	Wife	B. M. 1898	13. 1
68	Reids	Wife	B. M. 1900	13. 1
69	Reids	Wife	B. M. 1902	13. 1
70	Reids	Wife	B. M. 1904	13. 1
71	Reids	Wife	B. M. 1906	13. 1
72	Reids	Wife	B. M. 1908	13. 1
73	Reids	Wife	B. M. 1910	13. 1
74	Reids	Wife	B. M. 1912	13. 1
75	Reids	Wife	B. M. 1914	13. 1
76	Reids	Wife	B. M. 1916	13. 1
77	Reids	Wife	B. M. 1918	13. 1
78	Reids	Wife	B. M. 1920	13. 1
79	Reids	Wife	B. M. 1922	13. 1
80	Reids	Wife	B. M. 1924	13. 1
81	Reids	Wife	B. M. 1926	13. 1
82	Reids	Wife	B. M. 1928	13. 1
83	Reids	Wife	B. M. 1930	13. 1
84	Reids	Wife	B. M. 1932	13. 1
85	Reids	Wife	B. M. 1934	13. 1
86	Reids	Wife	B. M. 1936	13. 1
87	Reids	Wife	B. M. 1938	13. 1
88	Reids	Wife	B. M. 1940	13. 1
89	Reids	Wife	B. M. 1942	13. 1
90	Reids	Wife	B. M. 1944	13. 1
91	Reids	Wife	B. M. 1946	13. 1
92	Reids	Wife	B. M. 1948	13. 1
93	Reids	Wife	B. M. 1950	13. 1
94	Reids	Wife	B. M. 1952	13. 1
95	Reids	Wife	B. M. 1954	13. 1
96	Reids	Wife	B. M. 1956	13. 1
97	Reids	Wife	B. M. 1958	13. 1
98	Reids	Wife	B. M. 1960	13. 1
99	Reids	Wife	B. M. 1962	13. 1
100	Reids	Wife	B. M. 1964	13. 1
101	Reids	Wife	B. M. 1966	13. 1
102	Reids	Wife	B. M. 1968	13. 1
103	Reids	Wife	B. M. 1970	13. 1

<http://lod.cedar-project.nl/maps/>



Querying Semantic Web Census Data

← → ↻ 🏠 yasgui.org

pwn3d guys x +

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


```
1 PREFIX leri: <http://lod.cedar-project.nl:8888/linked-edit-rules/resource/>
2 PREFIX cedarterms: <http://bit.ly/cedar#>
3 PREFIX qb: <http://purl.org/linked-data/cube#>
4 PREFIX ler: <http://bit.ly/linked-edit-rules#>
5
6 SELECT * WHERE {
7   ?obs a qb:Observation.
8   ?measure a qb:MeasureProperty .
9   ?obs ?measure ?population .
10  FILTER (datatype(?population) != xsd:integer)
11 } LIMIT 1000
```

🔍 📄 📄 📄 📄 📄

Showing 1 to 50 of 1,000 entries (in 0.201 seconds)

obs	measure	population
1 http://id.insee.fr/demo/pop5/2010/observation/28302-2-025-21	http://rdf.insee.fr/meta/demo/mesure/pop2010	"0.0"^^xsd:float
2 http://id.insee.fr/demo/pop5/2010/observation/28302-2-025-22	http://rdf.insee.fr/meta/demo/mesure/pop2010	"0.0"^^xsd:float
3 http://id.insee.fr/demo/pop5/2010/observation/28302-2-025-24	http://rdf.insee.fr/meta/demo/mesure/pop2010	"5.12652"^^xsd:float
4 http://id.insee.fr/demo/pop5/2010/observation/28302-2-025-26	http://rdf.insee.fr/meta/demo/mesure/pop2010	"0.0"^^xsd:float
5 http://id.insee.fr/demo/pop5/2010/observation/28302-2-030-11	http://rdf.insee.fr/meta/demo/mesure/pop2010	"25.6326"^^xsd:float
6 http://id.insee.fr/demo/pop5/2010/observation/28302-2-030-12	http://rdf.insee.fr/meta/demo/mesure/pop2010	"0.0"^^xsd:float

Search:



URI	name	population
http://dbpedia.org/resource/Bia%C5%82a_G%C3%B3ra_Podd%C4%99bice_County	Biała Góra, Poddebice County	-1539607552
http://dbpedia.org/resource/Asian_people	Asian people	-964967296
http://dbpedia.org/resource/Asia	Asia	-415967296
http://dbpedia.org/resource/Phmstead	Phmstead	-99596434
http://dbpedia.org/resource/Signy_Island		

Negative Population



s	name	population
http://dbpedia.org/resource/Cotati%2C_California	Cotati, California	Cotatian
http://dbpedia.org/resource/New_York_City	New York City	New Yorker
http://dbpedia.org/resource/New_York_City	New York City	New Yorker
http://dbpedia.org/resource/Lima	Lima	Limeño/a

Weird Population Values



URI	URL
http://dbpedia.org/resource/Meleka	http://www.Meleka.co.uk
http://dbpedia.org/resource/University_of_Montana_School_of_Journalism	http://www.jour.umt.edu
http://dbpedia.org/resource/Santa_Cilia	http://www.santacilia.es
http://dbpedia.org/resource/Trick_Daddy	http://www.trickdaddy.com

Invalid URL's



Data retrieved on 2011-03-12 from <http://loc.openlinksw.com/sparql>

Introducing (Micro-) Edit Rules

```
1 dat1 : ageGroup %in% c('adult', 'child', 'elderly')
2 dat7 : maritalStatus %in% c('married', 'single', 'widowed')
3 num1 : 0 <= age
4 num2 : 0 < height
5 num3 : age <= 150
6 num4 : yearsMarried < age
7 cat5 : if(ageGroup == 'child') maritalStatus != 'married'
8 mix6 : if(age < yearsMarried + 17) !(maritalStatus %in% c('married', 'widowed'))
9 mix7 : if(ageGroup %in% c('adult', 'elderly') age >= 18
10 mix8 : if(ageGroup %in% c('child', 'elderly') & 18 <= age) age >= 65
11 mix9 : if(ageGroup %in% c('adult', 'child')) 65 > age
```

Listing 1.1: Examples of micro-edits in the R editrules package.

	age	ageGroup	height	maritalStatus	yearsMarried
#1	21	adult	6.0	single	-1
#2	2	child	3	married	0
#3	18	adult	5.7	married	20
#4	221	elderly	5	widowed	2
#5	34	child	-7	married	3

Table 1: Example dataset to be validated against obvious inconsistencies.

Introducing (Macro-) Edit Rules

	age	ageGroup	height	maritalStatus	yearsMarried
#1	21	adult	6.0	single	-1
#2	2	child	3	married	0
#3	18	adult	5.7	married	20
#4	221	elderly	5	widowed	2
#5	34	child	-7	married	3

Table 1: Example dataset to be validated against obvious inconsistencies.

Distribution methods

- Population count must be log-normal distributed

Aggregation methods

- Total population count should match the logical composition of records

Introducing Edit Rules

```
1 dat1 : ageGroup %in% c('adult', 'child', 'elderly')
2 dat7 : maritalStatus %in% c('married', 'single', 'widowed')
3 num1 : 0 <= age
4 num2 : 0 < height
5 num3 : age <= 150
6 num4 : yearsMarried < age
7 cat5 : if(ageGroup == 'child') maritalStatus != 'married'
8 mix6 : if(age < yearsMarried + 17) !(maritalStatus %in% c('married', 'widowed'))
9 mix7 : if(ageGroup %in% c('adult', 'elderly') age >= 18
10 mix8 : if(ageGroup %in% c('child', 'elderly') & 18 <= age) age >= 65
11 mix9 : if(ageGroup %in% c('adult', 'child')) 65 > age
```

Listing 1.1: Examples of micro-edits in the R editrules package.

Distribution methods

- Population count must be log-normal distributed

Aggregation methods

- Total population count should match the logical composition of records

Q: Do I have to write these for all my datasets?

A: YES



Share? Exchange? Reuse?

Requirements 1: Web friendly

- Published on and distributed over the Web
- Web structured data format
 - Machine readable
- Uniquely identifiable
 - Unambiguously dereferenceable

Requirements 2:

Edit rules as enriched SW rules

- Compatible with SW rule languages standards
 - Reuse existing work in formal languages
- Linked to constrained QB dimensions
 - Explicit about the dimensions the rules constrain
- Explicit scope
 - Explicit about the scope and the level (micro/macro)

Requirements 3:

QB (obvious) consistencies

- Integration with existing reasoning
 - Transparency and interoperability with current infrastructure
- Users free to choose which rules to run against which data cubes

Requirements 4:

Web standard reporting

- Provenance of the check-up process
 - Explicit and traceable annotation of the consistency checking (and enforcement) process
- Annotation of inconsistencies
 - Rules annotate inconsistent data points for expert correction

State of the art

	Web friendly	SW Rules	QB Consistency	Reporting
editrules, validate			✓	○
Eurostat's EVE			✓	
SDMX VTL			✓	
OWL 2	✓	✓	○	
SWRL	✓	✓	○	
SPARQL CONSTRUCT		○	○	
RIF-BLD	✓	✓	○	
SPIN	✓	✓	○	
Shape Expressions	✓	✓		
Resource Shapes	✓	✓		
OWLIM Profiles		✓	○	
TopBraid	✓	✓	○	○
Stardog	✓	✓	○	○
RDF Data Cube		✓	✓	

Table 14: Reviewed approaches according to their coverage of requirements by topic of Section 7.2. A circle (○) indicates that the approach covers some requirement in that topic, but not all; a check mark (✓) means that all requirements in the topic are met.

Linked Edit Rules

- A methodology to **publish, link, combine** and **execute *edit rules*** on the **Web as Linked Data...**
- ... to verify **consistency** of statistical datasets.

```
@prefix rule: <tag:stardog:api:rule> .
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> .
@prefix leri: <http://lod.cedar-project.nl:8888/linked-edit-rules/resource/> .
@prefix ler: <http://bit.ly/linked-edit-rules#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix eg: <http://example.org/ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# num1 : 0 <= age

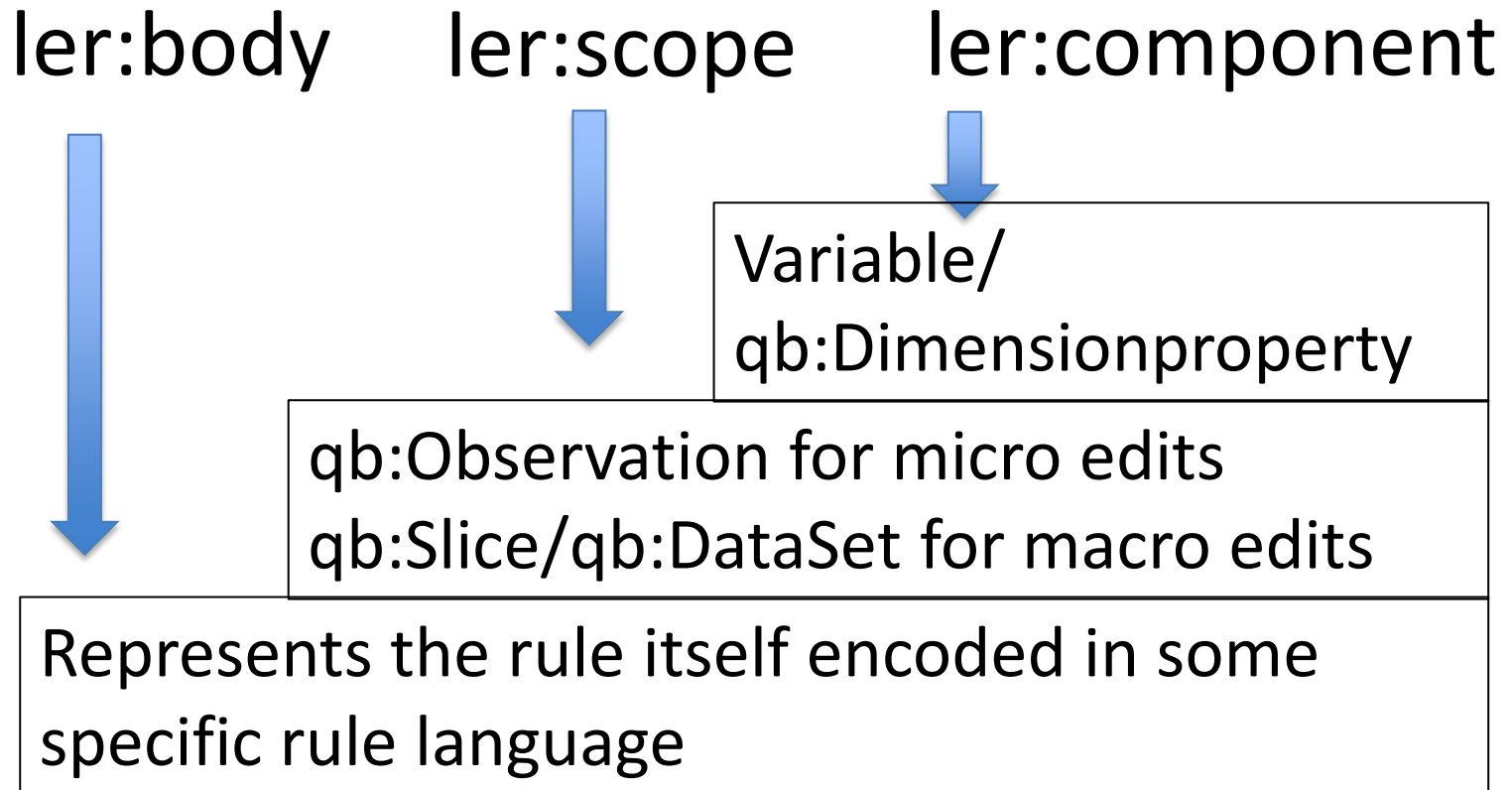
leri:num1 a rule:SPARQLRule, ler:EditRule ;
  rule:content """
    PREFIX leri: <http://lod.cedar-project.nl:8888/linked-edit-rules/resource/>
    PREFIX sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>
    PREFIX qb: <http://purl.org/linked-data/cube#>
    PREFIX ler: <http://bit.ly/linked-edit-rules#>
  IF {
    ?obs a qb:Observation .
    ?obs sdmx-dimension:age ?age .
    FILTER (?age < 0)
  }
  THEN {
    ?obs ler:inconsistentWith leri:num1 .
  }
  """ ;
  ler:scope qb:Observation ;
  ler:component sdmx-dimension:age ;
  rdfs:label "0 <= age"@en ;
  rdfs:comment "Age must be zero or a positive number"@en ;
  dc:creator <http://www.albertmeronyo.org/> ;
  dc:date "2015-01-08T16:03:40+01:00"^^xsd:dateTime
.
```



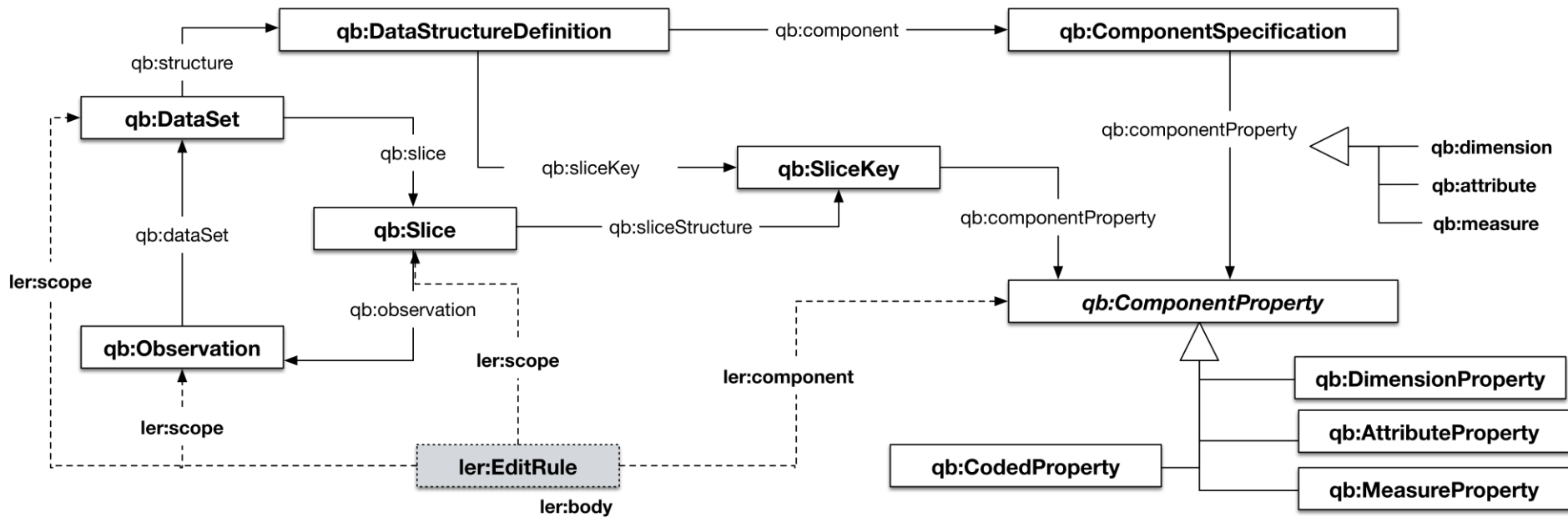
The **rule** and its **constrained dimensions** are **semantically connected**

Linked Edit Rules

- A `ler:EditRule` has three fundamental components:



LER as an extension of Data Cube Data model



Body of Micro-Edits

`if(ageGroup %in% c('child', 'elderly') & 18 <= age) age >= 65`

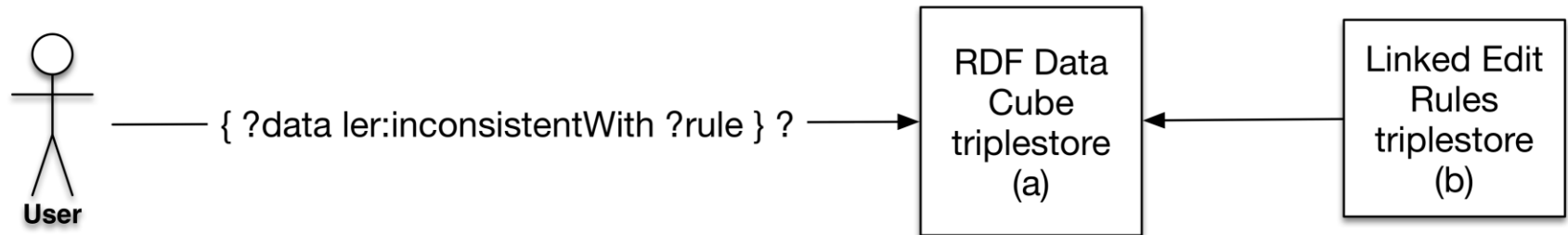
- Definite Horn Clauses, i.e. conjunctions of inequalities
- Convert variables and values to URIs and literals
 - Known dimension properties (sdmx-dimension:age)
 - String literals to equivalent values (sdmx-code:status-M)
 - Numeric values by RDF literals
- IF {
 ?obs a qb:Observation .
 ?obs eg:agegroup ?ageGroup .
 ?obs sdmx-dimension:age ?age .
 FILTER ((18 <= ?age && ?age < 65) && (?ageGroup = eg:child ||
 ?ageGroup = eg:elderly)) }
 THEN
 { ?obs ler:inconsistentWith leri:mix8 . }

Body of Macro-Edits

- $\text{Statistic}(t, P, S, c)$ as inequalities
 - t is a test (z.test normality test)
 - P are the parameters of t (e.g. Mean, variance)
 - S set of observations
 - c is the constrained dimension
- Example:
$$\text{statistic}(\text{z.test}, \{\mu, \sigma^2\}, \text{eg:all}, \text{eg:height}) > 0.05$$

Architecture

- User sends Query : which qb:Observation, qb:Slice or qb:DataSet is inconsistent with a specific ler:EditRule



Stardog Linked Micro-Edit Rule

```
# Micro-edit
leri:mix6 a rule:SPARQLRule, ler:EditRule;
  rule:content "" # PREFIX definitions
  IF {
    ?obs a qb:Observation.
    ?obs sdmx-dimension:civilStatus ?civilStatus.
    ?obs eg:yearsMarried ?yearsMarried.
    ?obs sdmx-dimension:age ?age.
    FILTER ((?age < ?yearsMarried + 17) && (?civilStatus = sdmx-code:status-M || ?
      civilStatus = sdmx-code:status-W))
  } THEN {
    ?obs ler:inconsistentWith leri:mix6.
  } "";
ler:scope qb:Observation;
ler:component sdmx-dimension:age, sdmx-dimension:civilStatus, eg:yearsMarried;
rdfs:label "if(age < yearsmarried + 17) !(status %in% c('married', 'widowed'))";
rdfs:comment "An underage can't be married nor widowed";
dc:creator <http://www.albertmeronyo.org/>;
dc:date "2015-01-08T16:03:40+01:00"^^xsd:dateTime.
```

Stardog Linked Macro-Edit Rule

```
# Macro-edit
leri:macro1 a rule:SPARQLRule, ler:EditRule;
  rule:content "" # PREFIX definitions
  IF {
    ?x a qb:Slice .
    ?x qb:sliceStructure eg:sliceByAdults .
    ?y a qb:Slice .
    ?y qb:sliceStructure eg:sliceByNonAdults .
    FILTER(stardog:R('wilcox.test', ?x, ?y, eg:height) <= 0.05)
```

SPARQL custom functions through R wrapper implemented as Stardog extension.

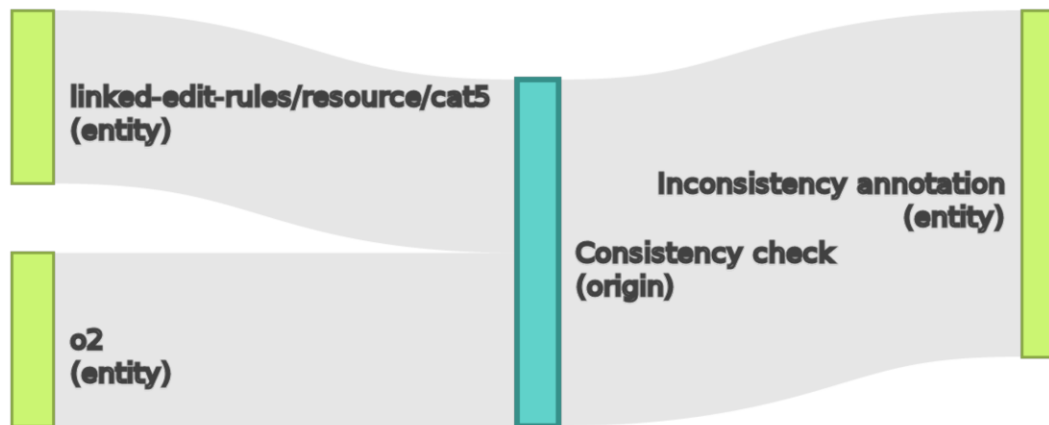
```
  ler:scope qb:Slice ;
  ler:component eg:height .
  rdfs:label "dist(X) != dist(Y), X heights of adults, Y heights of non-adults";
  rdfs:comment "Heights of adults and non-adults follow different distribs.";
  dc:creator <http://www.albertmeronyo.org/>;
  dc:date "2015-01-08T16:03:40+01:00"^^xsd:dateTime.
```

Triggering Rules in Stardog

```
1  # PREFIX definitions
2  INSERT { ?act a prov:Activity;
3          rdfs:label "Consistency check";
4          prov:wasAssociatedWith <http://stardog.com/>;
5          prov:startedAtTime ?now;
6          prov:used ?dp;
7          prov:used ?rule .
8  ?ann a oa:Annotation;
9          rdfs:label "Inconsistency annotation";
10         prov:wasGeneratedBy ?act;
11         prov:generatedAtTime ?now;
12         oa:hasBody ?body;
13         oa:hasTarget ?dp .
14  ?body a rdfs:Resource;
15         ler:inconsistentWith ?rule.
16  BIND (UUID() AS ?act, UUID() AS ?ann, UUID() AS ?
17         body, now() AS ?now)
17 } WHERE { ?dp ler:inconsistentWith ?rule . }
```

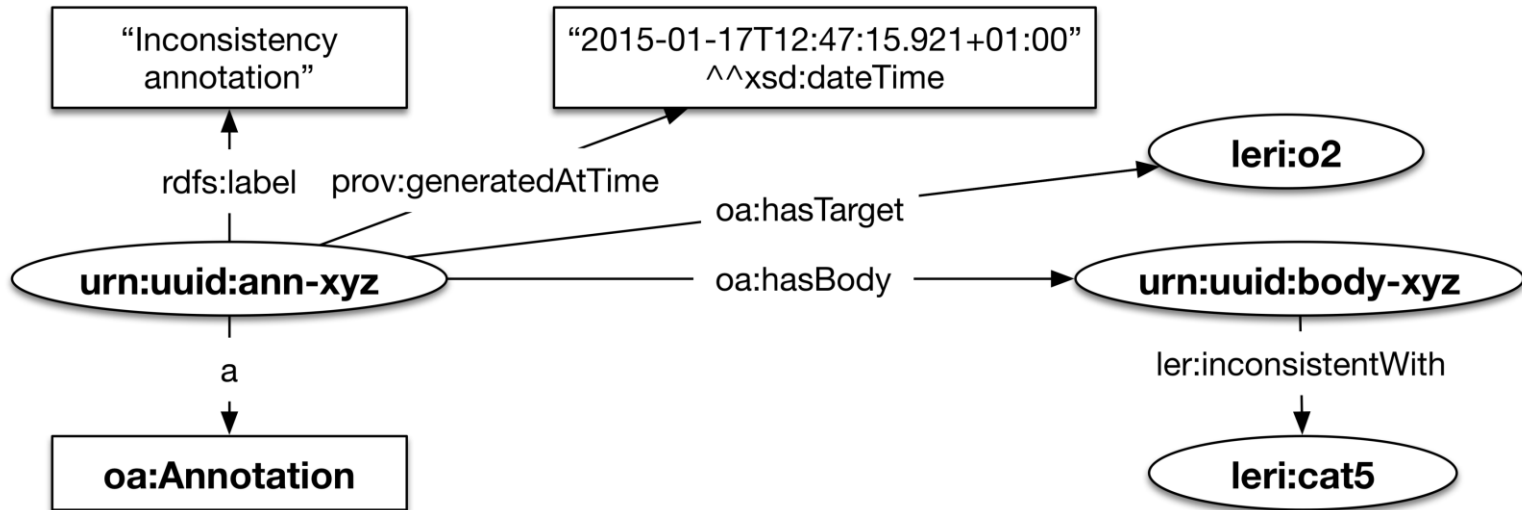
Listing 7.4: SPARQL INSERT that triggers Linked Micro- and Macro-Edit Rules in Stardog. For each inferred inconsistency, PROV provenance and OA annotations are generated.

Provenance



(a) PROV-O-Viz [81] diagram showing a consistency check (activity) using observation `1eri:o2` (entity) and rule `1eri:cat5` (entity) to produce an inconsistency annotation (entity).

Annotation



(b) Generated OA annotation of a detected inconsistency in a `qb:Observation`, linking to the inconsistent observation `leri:o2` as target, and describing the inconsistency (w.r.t. rule `leri:cat5`) as body.

Simple stress testing

- Datasets
 - Toy data from previous slide
 - Synthetic Dataset
- Result: correct identification of expected inconsistencies (with minimal runtime 0.5%)

Showcasing in two real datasets

- Datasets
 - CEDAR usecase
- Results:
 - 4.5% have no occupation positions
 - 0.56% count population have non-integer numbers
 - 0.02% have negative numbers

Take-home message!

Linked Data Rules as

- A framework and a data model to express Edit Rules as Linked Edit Rules (LER).
 - Allowing to **Link, retrieve, reuse, combine** and **execute** Edit Rules on the Web to check **quality** and **consistency** of RDF Data Cubes
- An automatic consistency checker that
 - Finds inconsistencies and generates **provenance reports** and **annotations**

Questions? Ask Albert: albert.merono@vu.nl