

Sviluppo di un Algoritmo per la Generazione e la Verifica Automatica di Vincoli su Grafi RDF

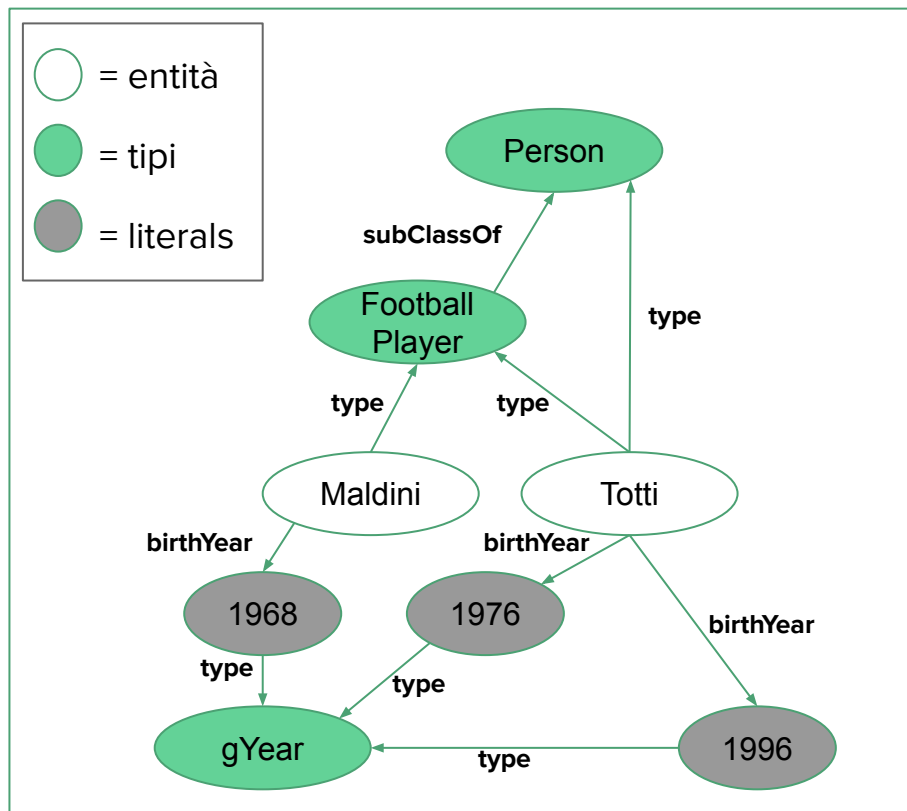


Christian Bernasconi #816423

Relatore: Prof. Andrea Maurino

Co-relatore: Ing. Blerina Spahiu, PhD

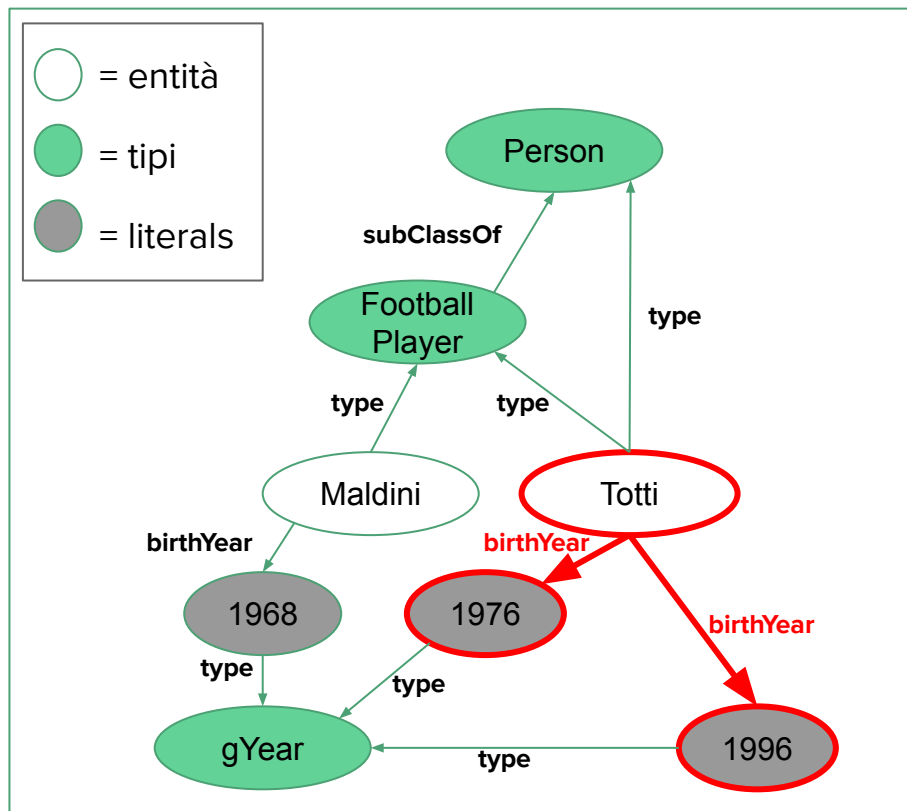
Contesto



Linked Data:

- RDF
- Tripla: soggetto predicato oggetto
- Entità, Tipi e Literals
- Identificazione univoca con URI

Contesto



Linked Data:

- RDF
- Tripla: soggetto predicato oggetto
- Entità, Tipi e Literals
- Identificazione univoca con URI

Errori cardinalità:

- Functional Properties
- Inconsistenze

Obiettivo dello stage

Fornire uno strumento per l'individuazione di potenziali errori di cardinalità garantendo:

- generazione automatica shapes
- verifica automatica vincoli
- applicabilità su qualsiasi dataset RDF
- nessuna richiesta di competenze tecniche

Punto di partenza: Abstat

Strumento di supporto per l'esplorazione di grandi dataset tramite profilazione

summary

pattern

statistiche

Subject Type (occurrences)	Predicate (occurrences)	Object Type (occurrences)	Frequency	Instances	Max Subjs-Obj	Avg Subjs-Obj	Min Subjs-Obj	Max Subj-Objs	Avg Subj-Objs	Min Subj-Objs
subject	predicate	object								
foaf:Person (1179233)	foaf:name (5514518)	rdfs:Literal (13460122)	3059104	-	185	1	1	32	3	1
foaf:Person (1179233)	dce:description (2361985)	rdfs:Literal (13460122)	2345088	-	32304	6	1	6	2	1
foaf:Person (1179233)	dbo:birthDate (1891809)	xmls:date (3009470)	1615538	-	153	17	1	6	2	1
dbo:Person (611330)	foaf:name (5514518)	rdfs:Literal (13460122)	1323375	-	49	1	1	32	2	1
dbo:Person (611330)	dce:description (2361985)	rdfs:Literal (13460122)	1173535	-	11647	5	1	6	2	1

Generazione shapes: ShaclGenerator

```
dbo:Person a sh:NodeShape;  
  sh:targetClass dbo:Person ;  
  sh:property [  
    sh:message "Errore cardinalita diretta" ;  
    sh:path dbo:residence ;  
    sh:qualifiedValueShape [ sh:class dbo:City ; ];  
    sh:qualifiedMinCount 1 ;  
    sh:qualifiedMaxCount 2 ;  
    sh:severity sh:Warning ;  
    sh:property [  
      sh:path [ sh:inversePath dbo:residence ; ];  
      sh:message "Errore cardinalita inversa" ;  
      sh:class dbo:Person ;  
      sh:minCount 1 ;  
      sh:maxCount 12 ;  
      sh:severity sh:Warning ; ] ]
```

- linguaggio SHACL

Generazione shapes: ShaclGenerator

```
dbo:Person a sh:NodeShape;  
  sh:targetClass dbo:Person ;  
  sh:property [  
    sh:message "Errore cardinalita diretta" ;  
    sh:path dbo:residence ;  
    sh:qualifiedValueShape [ sh:class dbo:City ; ];  
    sh:qualifiedMinCount 1 ;  
    sh:qualifiedMaxCount 2 ;  
    sh:severity sh:Warning ;  
  ]  
  sh:property [  
    sh:path [ sh:inversePath dbo:residence ; ];  
    sh:message "Errore cardinalita inversa" ;  
    sh:class dbo:Person ;  
    sh:minCount 1 ;  
    sh:maxCount 12 ;  
    sh:severity sh:Warning ; ] ]
```

- linguaggio SHACL

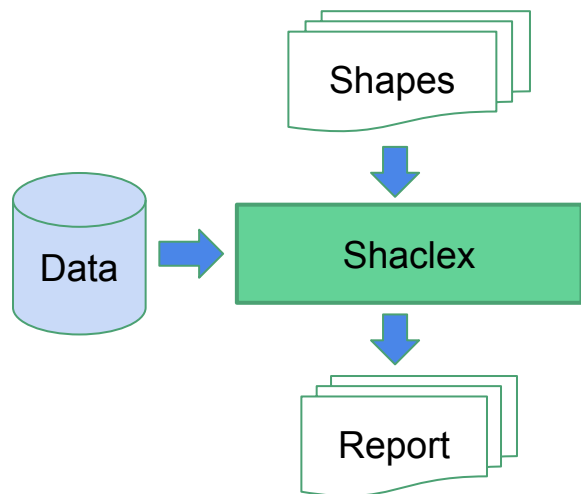
Generazione shapes: ShaclGenerator

```
dbo:Person a sh:NodeShape;  
sh:targetClass dbo:Person ;  
sh:property [  
  sh:message "Errore cardinalita diretta" ;  
  sh:path dbo:residence ;  
  sh:qualifiedValueShape [ sh:class dbo:City ; ];  
  sh:qualifiedMinCount 1 ;  
  sh:qualifiedMaxCount 2 ;  
  sh:severity sh:Warning ;  
  sh:property [  
    sh:path [ sh:inversePath dbo:residence ; ];  
    sh:message "Errore cardinalita inversa" ;  
    sh:class dbo:Person ;  
    sh:minCount 1 ;  
    sh:maxCount 12 ;  
    sh:severity sh:Warning ; ] ]
```

- linguaggio SHACL
- relazione shape-pattern

Subject Type (occurrences)	Predicate (occurrences)	Object Type (occurrences)	Frequency	Instances	Max Subjs-Obj	(x2) Avg Subjs-Obj	Min Subjs-Obj	Max Subj-Objs	(x2) Avg Subj-Objs	Min Subj-Objs
dbo:Person (611330)	dbo:residence (62168)	dbo:City (20943)	7395	22045	760	6	1	9	1	1

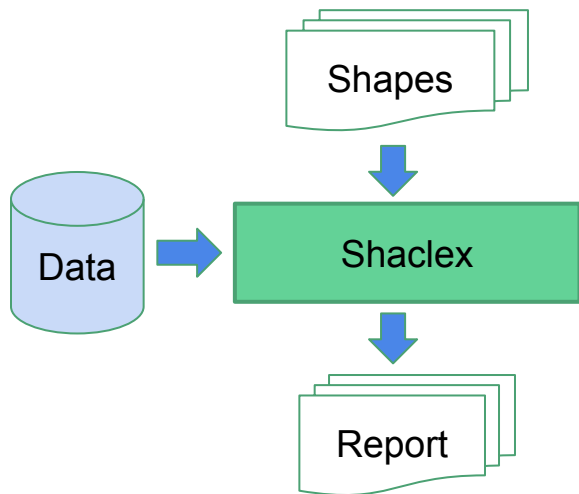
Verifica vincoli: Shaclex



Validatore SHACL:

- verifica presenza violazioni
- raccoglie errori in report

Verifica vincoli: Shaclex



Validatore SHACL:

- verifica presenza violazioni
- raccoglie errori in report

Problemi emersi:

- validazione via endpoint
- esigenze pattern Abstat
- bug risorse di tipo *date*

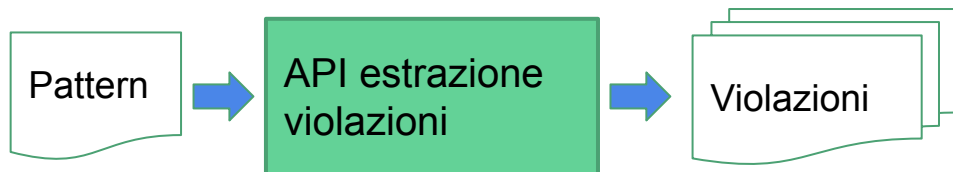
Verifica vincoli: API estrazione triple

In sostituzione al validatore sono state implementate delle RESTful API per l'individuazione degli errori di cardinalità

- validazione **on-demand** del pattern
- traduzione vincoli in **SPARQL queries**
- raccolta errori in un **JSON**
- verifica divisa in **due fasi**

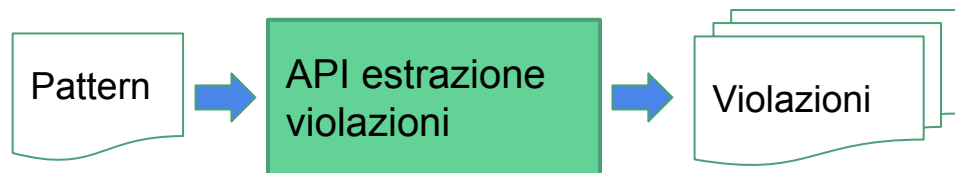
Verifica vincoli: API estrazione triple

Fase 1: estrazione violazioni

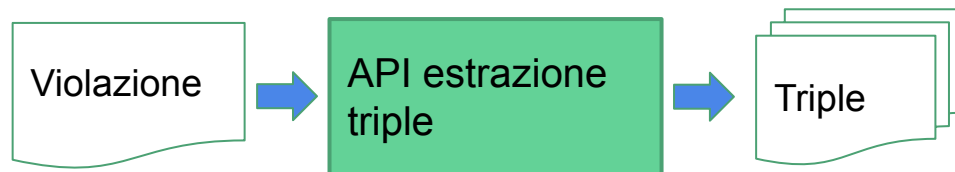


Verifica vincoli: API estrazione triple

Fase 1: estrazione violazioni

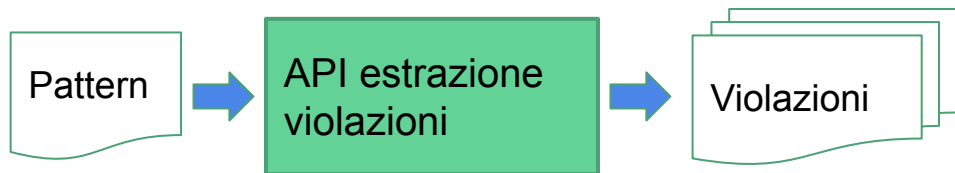


Fase 2: estrazione triple violazione

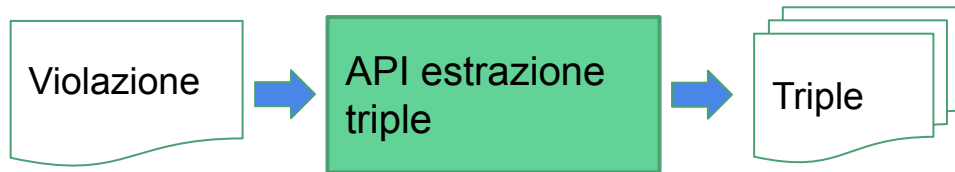


Verifica vincoli: API estrazione triple

Fase 1: estrazione violazioni



Fase 2: estrazione triple violazione

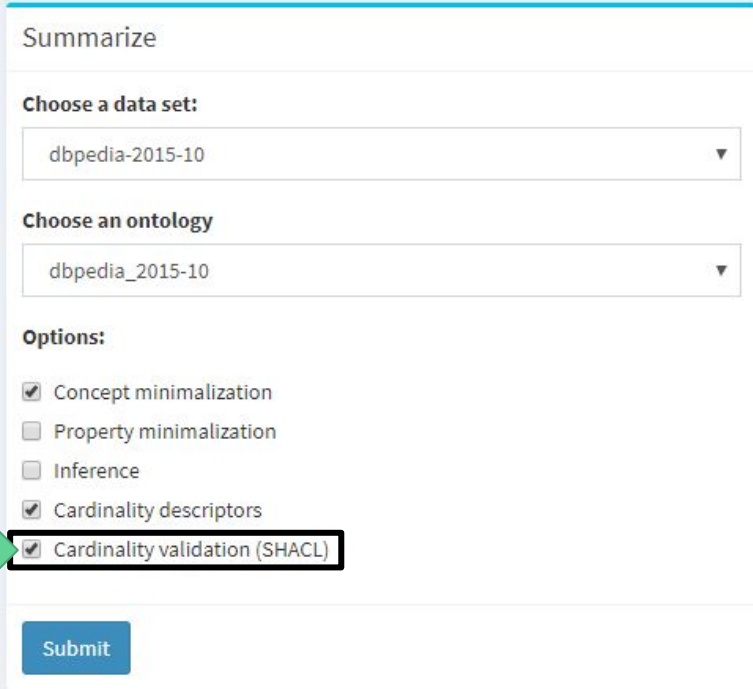


Vantaggi:

- query meno complesse
- JSON più gestibili lato client
- migliora esperienza utente

Esempio di utilizzo: indagine DBpedia 2015

Profilazione e generazione shapes
per il dataset di DBpedia nella
release 2015-10



Summarize

Choose a data set:

dbpedia-2015-10 ▼

Choose an ontology

dbpedia_2015-10 ▼






Options:

- ☒ Concept minimization
- ☐ Property minimization
- ☐ Inference
- ☒ Cardinality descriptors
- ☒ Cardinality validation (SHACL)

Submit






Esempio di utilizzo: indagine DBpedia 2015

Esplorazione dei pattern

Subject Type (occurrences)	Predicate (occurrences)	Object Type (occurrences)	Frequency	Instances	Max Subjs-Obj	Avg Subjs-Obj	Min Subjs-Obj	Max Subj-Objs	Avg Subj-Objs	Min Subj-Objs	Severity
<input type="text" value="subject"/>	<input type="text" value="predicate"/>	<input type="text" value="object"/>									
dbo:Book (34591)	foaf:name (5514518)	rdfs:Literal (13460122)	37731	-	6	1	1	50	1	1	 details
dbo:Book (34591)	dbo:author (57260)	foaf:Person (1179233)	30105	-	128	3	1	16	1	1	 details
dbo:Book (34591)	dbo:literaryGenre (29471)	owl:Thing (95380)	28029	-	2200	26	1	9	1	1	 details
dbo:Book (34591)	dbo:isbn (26633)	rdfs:Literal (13460122)	26629	-	1850	1	1	33	1	1	 details
dbo:Book (34591)	dbo:numberOfPages (24842)	xmls:positiveInteger (42789)	24816	-	572	22	1	17	1	1	 details


Esempio di utilizzo: indagine DBpedia 2015

Esplorazione dei pattern

Subject Type (occurrences)	Predicate (occurrences)	Object Type (occurrences)	Frequency	Instances	Max Subjs-Obj	Avg Subjs-Obj	Min Subjs-Obj	Max Subj-Objs	Avg Subj-Objs	Min Subj-Objs	Severity
<input type="text" value="subject"/>	<input type="text" value="predicate"/>	<input type="text" value="object"/>									
dbo:Book (34591)	foaf:name (5514518)	rdfs:Literal (13460122)	37731	-	6	1	1	50	1	1	 details
dbo:Book (34591)	dbo:author (57260)	foaf:Person (1179233)	30105	-	128	3	1	16	1	1	 details
dbo:Book (34591)	dbo:literaryGenre (29471)	owl:Thing (95380)	28029	-	2200	26	1	9	1	1	 details
dbo:Book (34591)	dbo:isbn (26633)	rdfs:Literal (13460122)	26629	-	1850	1	1	33	1	1	 details
dbo:Book (34591)	dbo:numberOfPages (24842)	xmls:positiveInteger (42789)	24816	-	572	22	1	17	1	1	 details

Esempio di utilizzo: indagine DBpedia 2015

Ispezione di un pattern (Fase 1 API)

Subject Type (occurrences)	Predicate (occurrences)	Object Type (occurrences)	Frequency	Instances	Max Subjs-Obj	Avg Subjs-Obj	Min Subjs-Obj	Max Subj- Objs	Avg Subj- Objs	Min Subj- Objs	Severity
dbo:Book (34591)	dbo:isbn (26633)	rdfs:Literal (13460122)	26629	-	1850	1	1	33	1	1	

Inspect	# Distinct Subjects	Predicate	Object
<input checked="" type="radio"/>	1850	dbo:isbn	NA
<input type="radio"/>	361	dbo:isbn	N/A
<input type="radio"/>	49	dbo:isbn	n/a
<input type="radio"/>	44	dbo:isbn	ISBN
<input type="radio"/>	39	dbo:isbn	Paperback:

5 matches found - get more






Inspect	Subject	Predicate	# Distinct Objects
<input checked="" type="radio"/>	dbr:List_of_Little_Miss_characters	dbo:isbn	33
<input type="radio"/>	dbr:List_of_Mr_Men	dbo:isbn	26
<input type="radio"/>	dbr:List_of_Transformers_books	dbo:isbn	22
<input type="radio"/>	dbr:Cthulhu_Mythos_anthology	dbo:isbn	18
<input type="radio"/>	dbr:Kitty_Norville	dbo:isbn	12

5 matches found - get more

Esempio di utilizzo: indagine DBpedia 2015

Ispezione triple coinvolte (Fase 2 API)

Possible cardinality violation (subject-wise)

Inspect	# Distinct Subjects	Predicate	Object
	1850	dbo:isbn	NA
	361	dbo:isbn	N/A
	49	dbo:isbn	n/a
	44	dbo:isbn	ISBN
	39	dbo:isbn	Paperback:

5 matches found - get more






Triples

Subject	Predicate	Object
dbr:A_Curtain_of_Green	dbo:isbn	NA
dbr:A_Mule_for_the_Marquesa	dbo:isbn	NA
dbr:A_Question_of_Upbringing	dbo:isbn	NA
dbr:Arthur_Mervyn	dbo:isbn	NA
dbr:Auriol_(novel)	dbo:isbn	NA
dbr:Call_It_Sleep	dbo:isbn	NA
dbr:Canary_in_a_Cathouse	dbo:isbn	NA
dbr:Cat_and_Mouse_(novella)	dbo:isbn	NA
dbr:Dangling_Man	dbo:isbn	NA
dbr:Dead_Fingers_Talk	dbo:isbn	NA






10 triples found - get more

Esempio di utilizzo: confronto versioni

2015






Inspect	# Distinct Subjects	Predicate	Object
	1850	dbo:isbn	NA
	361	dbo:isbn	N/A
	49	dbo:isbn	n/a
	44	dbo:isbn	ISBN
	39	dbo:isbn	Paperback:
5 matches found - get more			

2016






Inspect	# Distinct Subjects	Predicate	Object
	5	dbo:isbn	1-84255-170-1
	3	dbo:isbn	0-7653-0330-2
	3	dbo:isbn	0-224-01924-4
	3	dbo:isbn	0-345-45644-0
	3	dbo:isbn	978-0-330-42389-2
5 matches found - get more			

Esempio di utilizzo: confronto versioni

2015






Inspect	# Distinct Subjects	Predicate	Object
	1850	dbo:isbn	NA
	361	dbo:isbn	N/A
	49	dbo:isbn	n/a
	44	dbo:isbn	ISBN
	39	dbo:isbn	Paperback:
5 matches found - get more			

2016






Inspect	# Distinct Subjects	Predicate	Object
	5	dbo:isbn	1-84255-170-1
	3	dbo:isbn	0-7653-0330-2
	3	dbo:isbn	0-224-01924-4
	3	dbo:isbn	0-345-45644-0
	3	dbo:isbn	978-0-330-42389-2
5 matches found - get more			

Esempio di utilizzo: confronto versioni

2015

Inspect	# Distinct Subjects	Predicate	Object
	1850	dbo:isbn	NA
	361	dbo:isbn	N/A
	49	dbo:isbn	n/a
	44	dbo:isbn	ISBN
	39	dbo:isbn	Paperback:
5 matches found - get more			

2016

Inspect	# Distinct Subjects	Predicate	Object
	5	dbo:isbn	1-84255-170-1
	3	dbo:isbn	0-7653-0330-2
	3	dbo:isbn	0-224-01924-4
	3	dbo:isbn	0-345-45644-0
	3	dbo:isbn	978-0-330-42389-2
5 matches found - get more			

Esempio di utilizzo: analisi correttezza

Sulla base del lavoro svolto è possibile proporre per il futuro una metodologia di analisi per calcolare degli indici di correttezza

	AKP	vers.	stato	freq.	cardinalità (M-a-m) ¹	sogg. violati	triple valide	errori (c-n-e) ²
1	dbo:Train, dbo:weight, xmls:double	2015 2016	ko ko	502 568	4 - 1 - 1 4 - 1 - 1	1.01% 1.07%	98.40% 98.41%	0 - 1 - 5
2	dbo:Continent, dbo:populationTotal, xmls:nonNegativeInteger	2015 2016	ko ok	23 17	6 - 1 - 1 1 - 1 - 1	5.5% 0%	78.2% 100%	1 - 0 - 0
3	dbo:BusinessPerson, dbo:birthYear, xmls:gYear	2015 2016	ko ok	144 89	2 - 1 - 1 1 - 1 - 1	0.7% 0%	99.3% 100%	1 - 0 - 0
4	dbo:Automobile, dbo:wheelbase, xmls:double	2015 2016	ok ko	5235 5442	1 - 1 - 1 5 - 1 - 1	0% 0.16%	100% 99.7%	/ - 9 - /

¹ M=massima; a=media; m=minima

² c=corretti; n=nuovi; e=ereditati

Summary stats
+
API estrazione
+
Controllo incrociato

Conclusioni e sviluppi futuri

Conclusioni:

- strumento semplice ed efficace
- passaggi di versione dataset non sempre positivi

Sviluppi futuri:

- machine-learning per previsione cardinalità
- calcolo di indici di correttezza

Grazie per l'attenzione
