Università degli Studi di Milano – Bicocca
Dipartimento di Informatica Sistemistica e Comunicazione

Towards Improving the Quality of Knowledge Graphs with Data-driven Ontology Patterns and SHACL

Blerina Spahiu, Andrea Maurino, Matteo Palmonari

spahiu|pamonari|maurino@disco.unimib.it

INSID&S Lab
Interaction and Semantics
for Innovation with Data & Services

@InsideLaBicocca
@blerinaspahiu

Outline

- Motivation
- Main Intuition
- SHACL
- Data-driven Ontology Patterns & KG profiles
- Actual Content vs. Desired Content
- SHACL Generation and Validation Methodology
- Conclusions and Future Work

Motivation: Quality of Knowledge Graphs & SHACL

- Understanding the content and evaluating the quality of data sets is challenging
- Many datasets extracted from semi-structured information
- Quality may change in different versions of the same data set
  - Check errors across different versions of data sets still in use
- Looking at the ontology is not enough
  - Ontologies may be large and underspecified
  - DBpedia 2015-04: 2795 properties, domain not specified for 259 properties, range not specified for 187 properties
  - No information about the usage
- SHACL to validate constraints
  - How to design SHACL profiles?

Main Intuition

Assist SHACL-based data validation using Knowledge Graphs (KG) profiles

KG Profile

	subject type (occurrences)	predicate (occurrences)	object type (occurrences)	frequency	instances	Max subjs-obj	Avg subjs-obj	Min subjs-obj	Max subj-objs
filter	dbo:Company	predicate	object						
	dbo:Company (51898)	DTP foaf:name (5002938)	rdfs:Literal (12948466)	52391	64155	200	1	1	16
	dbo:Company (51898)	OP foaf:homepage (362016)	owl:Thing	42861	52699	27	1	1	11
	dbo:Company (51898)	OP dbo:industry (50100)	owl:Thing	42408	50093	1750	10	1	15
	dbo:Company (51898)	DTP dbo:foundingYear (87412)	xmls:gYear (3469462)	42131	49085	1159	99	1	8

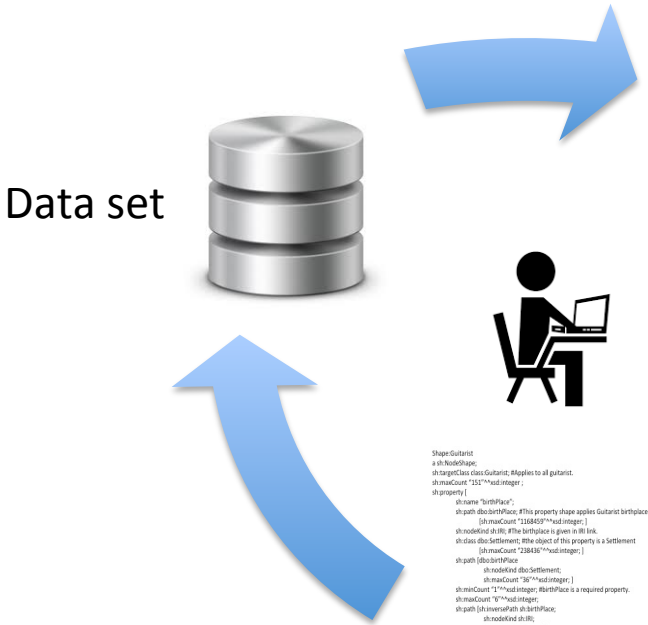
```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;       #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #BirthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

Main Intuition

Assist SHACL-based data validation using Knowledge Graphs (KG) profiles



KG Profile

	subject type (occurrences)	predicate (occurrences)	object type (occurrences)	frequency	instances	Max subjs-obj	Avg subjs-obj	Min subjs-obj	Max subj-objs
filter	dbo:Company	predicate	object						
	dbo:Company (51898)	DTP foaf:name (5002938)	rdfs:Literal (12948466)	52391	64155	200	1	1	16
	dbo:Company (51898)	OP foaf:homepage (362016)	owl:Thing	42861	52699	27	1	1	11
	dbo:Company (51898)	OP dbo:industry (50100)	owl:Thing	42408	50093	1750	10	1	15
	dbo:Company (51898)	DTP dbo:foundingYear (87412)	xmls:gYear (3469462)	42131	49085	1159	99	1	8

Tools providing KG profiles based on schema patterns

- ABSTAT
- LOUPE

- LODSTAT
- SCHEMEX
- …

Main Intuition

Assist SHACL-based data validation using Knowledge Graphs (KG) profiles

Data set

KG Profile

	subject type (occurrences)	predicate (occurrences)	object type (occurrences)	frequency	instances	Max subjs-obj	Avg subjs-obj	Min subjs-obj	Max subj-objs
filter	dbo:Company	predicate	object						
	dbo:Company (51898)	DTP foaf:name (5002938)	rdfs:Literal (12948466)	52391	64155	200	1	1	16
	dbo:Company (51898)	OP foaf:homepage (362016)	owl:Thing	42861	52699	27	1	1	11
	dbo:Company (51898)	OP dbo:industry (50100)	owl:Thing	42408	50093	1750	10	1	15
	dbo:Company (51898)	DTP dbo:foundingYear (87412)	xmls:gYear (3469462)	42131	49085	1159	99	1	8

SHACL Profile

Possible approaches

- Manual
- Heuristic
- Automatic

Tools providing KG profiles based on schema patterns

- ABSTAT
- LOUPE

- LODSTAT
- SCHEMEX
- …

Data-driven Ontology Patterns & KG profiles

ABSTAT profiles = data-driven ontology patterns + statistics:

- Data-driven ontology patterns: (minimal type) schema patterns, i.e., (most specific) patterns extracted from data
- Statistics: occurrence, frequency, instances, cardinality descriptors

	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
filter	dbo:Company	predicate	object								
	dbo:Company (51898)	DTP foaf:name (5002938)	rdfs:Literal (12948466)	52391	64155	200	1	1	16	1	1
	dbo:Company (51898)	OP foaf:homepage (362016)	owl:Thing	42861	52699	27	1	1	11	1	1
	dbo:Company (51898)	OP dbo:industry (50100)	owl:Thing	42408	50093	1750	10	1	15	1	1
	dbo:Company (51898)	DTP dbo:foundingYear (87412)	xmls:gYear (3469462)	42131	49085	1159	99	1	8	1	1

Data-driven Ontology Patterns & KG profiles

ABSTAT profiles = data-driven ontology patterns + statistics:

- Data-driven ontology patterns: (minimal type) schema patterns, i.e., (most specific) patterns extracted from data
- Statistics: occurrence, frequency, instances, cardinality descriptors

Schema patterns: there exist entities that have Company as *minimal type, which* are linked to literals that have gYear as *minimal type* by the property foundingYear

	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
filter	dbo:Company	predicate	object								
	dbo:Company (51898)	DTP foaf:name (5002938)	rdfs:Literal (12948466)	52391	64155	200	1	1	16	1	1
	dbo:Company (51898)	OP foaf:homepage (362016)	owl:Thing	42861	52699	27	1	1	11	1	1
	dbo:Company (51898)	OP dbo:industry (50100)	owl:Thing	42408	50093	1750	10	1	15	1	1
	dbo:Company (51898)	DTP dbo:foundingYear (87412)	xmls:gYear (3469462)	42131	49085	1159	99	1	8	1	1

Data-driven Ontology Patterns & KG profiles

ABSTAT profiles = data-driven ontology patterns + statistics:

- Data-driven ontology patterns: (minimal type) schema patterns, i.e., (most specific) patterns extracted from data
- Statistics: occurrence, frequency, instances, cardinality descriptors

Schema patterns: there exist entities that have Company as *minimal type, which* are linked to literals that have gYear as *minimal type* by the property foundingYear

	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
filter	dbo:Company	predicate	object								
	dbo:Company (51898)	DTP foaf:name (5002938)	rdfs:Literal (12948466)	52391	64155	200	1	1	16	1	1
	dbo:Company (51898)	OP foaf:homepage (362016)	owl:Thing	42861	52699	27	1	1	11	1	1
	dbo:Company (51898)	OP dbo:industry (50100)	owl:Thing	42408	50093	1750	10	1	15	1	1
	dbo:Company (51898)	DTP dbo:foundingYear (87412)	xmls:gYear (3469462)	42131	49085	1159	99	1	8	1	1

Occurrence of types and properties

Data-driven Ontology Patterns & KG profiles

ABSTAT profiles = data-driven ontology patterns + statistics:

- Data-driven ontology patterns: (minimal type) schema patterns, i.e., (most specific) patterns extracted from data
- Statistics: occurrence, frequency, instances, cardinality descriptors

Schema patterns: there exist entities that have Company as *minimal type, which* are linked to literals that have gYear as *minimal type* by the property foundingYear

Frequency and instances: how many times this pattern occurs as minimal type pattern. Instances count considers pattern inference

	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
filter	dbo:Company	predicate	object								
	dbo:Company (51898)	DTP foaf:name (5002938)	rdfs:Literal (12948466)	52391	64155	200	1	1	16	1	1
	dbo:Company (51898)	OP foaf:homepage (362016)	owl:Thing	42861	52699	27	1	1	11	1	1
	dbo:Company (51898)	OP dbo:industry (50100)	owl:Thing	42408	50093	1750	10	1	15	1	1
	dbo:Company (51898)	DTP dbo:foundingYear (87412)	xmls:gYear (3469462)	42131	49085	1159	99	1	8	1	1

Occurrence of types and properties

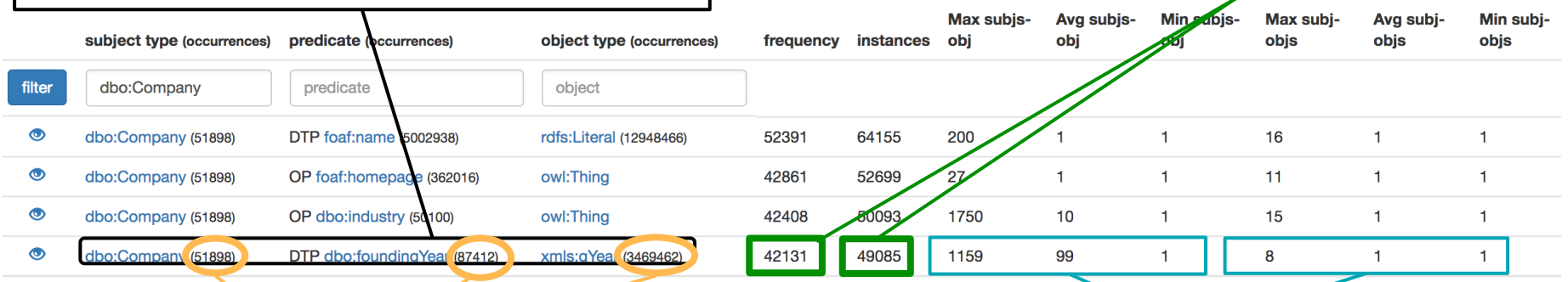
Data-driven Ontology Patterns & KG profiles

ABSTAT profiles = data-driven ontology patterns + statistics:

- Data-driven ontology patterns: (minimal type) schema patterns, i.e., (most specific) patterns extracted from data
- Statistics: occurrence, frequency, instances, cardinality descriptors

Schema patterns: there exist entities that have Company as *minimal type, which* are linked to literals that have gYear as *minimal type* by the property foundingYear

Frequency and instances: how many times this pattern occurs as minimal type pattern. Instances count considers pattern inference



	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
filter	dbo:Company	predicate	object								
	dbo:Company (51898)	DTP foaf:name (5002938)	rdfs:Literal (12948466)	52391	64155	200	1	1	16	1	1
	dbo:Company (51898)	OP foaf:homepage (362016)	owl:Thing	42861	52699	27	1	1	11	1	1
	dbo:Company (51898)	OP dbo:industry (50100)	owl:Thing	42408	50093	1750	10	1	15	1	1
	dbo:Company (51898)	DTP dbo:foundingYear (87412)	xmls:gYear (3469462)	42131	49085	1159	99	1	8	1	1

Occurrence of types and properties

Cardinality descriptors: max/avg/min number of different subjects associated with a same object (and vice versa)

SHACL

- Shapes Constraint Language (SHACL) is a W3C recommendation language for defining constraints on RDF graphs
- A SHACL processor has two inputs:
  - A ***data graph*** that contains the RDF data
  - A ***shapes graph*** that contains the shapes
- Two types of shapes:
  - ***Node shape*** that declare constraints directly on a node e.g., node kind.
  - ***Property shape*** that declare constraints on the values associated with a node through a path e.g., cardinality.
- The validation report produced by SHACL contains three different severity levels; ***Violation***, ***Warning*** and ***Info***.
- SHACL is divided into:
  - ***SHACL Core*** which describes a core RDF vocabulary
  - ***SHACL-SPARQL*** describes an extension mechanism in terms of SPARQL

SHACL Cardinality Constraints

- Cardinality constraint for the property schema:email for the resource of Bob

```
dbo:Person
        a sh:NodeShape ;
        sh:targetNode dbr:Bob ;
        sh:property [
                sh:path schema:email ;
                sh:minCardinality 1;
                sh:maxCardinality 2;
        ] .
```

Actual Content vs. Desired Content

Data set

```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;        #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #birthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;        #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #birthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

Actual Content vs. Desired Content

Data set

```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;        #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #birthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

Describing what ***is*** in the data set

```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;        #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #birthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

Actual Content vs. Desired Content

Data set

```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;       #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #birthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

Describing what ***is*** in the data set

```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;       #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #birthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

Describing what ***should be*** in the data set

Actual Content vs. Desired Content

Data set

```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;       #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #birthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

```
Shape:Guitarist
a sh:NodeShape;
sh:targetClass class:Guitarist; #Applies to all guitarist.
sh:maxCount "151"^^xsd:integer ;
sh:property [
        sh:name "birthPlace";
        sh:path dbo:birthPlace; #This property shape applies Guitarist birthplace.
                [sh:maxCount "1168459"^^xsd:integer; ]
        sh:nodeKind sh:IRI; #The birthplace is given in IRI link.
        sh:class dbo:Settlement; #the object of this property is a Settlement
                [sh:maxCount "238436"^^xsd:integer; ]
        sh:path [dbo:birthPlace
                sh:nodeKind dbo:Settlement;
                sh:maxCount "36"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthPlace is a required property.
        sh:maxCount "6"^^xsd:integer;
        sh:path [sh:inversePath sh:birthPlace;
                sh:nodeKind sh:IRI;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ]
        ];

sh:property [
        sh:name "birthDate";
        sh:path    dbo:birthDate;       #This property shape applies to guitarist's birthday.
                [sh:maxCount "1101418"^^xsd:integer; ]
        sh:datatype xmls:date; #birthDate is a date;
                [sh:maxCount "1884250"^^xsd:integer; ]
        sh:path [dbo:birthDate
                sh:datatype xmls:date;
                sh:maxCount "88"^^xsd:integer; ]
        sh:minCount "1"^^xsd:integer; #birthDate is a required property.
        sh:maxCount "1"^^xsd:integer;
        sh:path [ sh:inversePath sh:birthDate;
                sh:datatype xmls:date;
                sh:minCount "1"^^xsd:integer;
                sh:maxCount "2"^^xsd:integer; ];
        ];
```

SHACL Profile

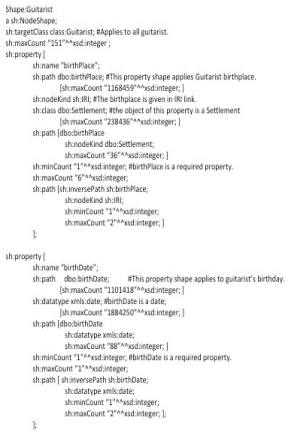
Describing what ***is*** in the data set

Describing what ***should be*** in the data set

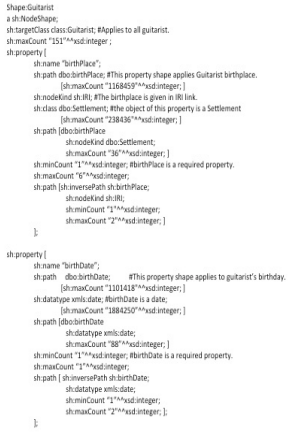
KGs Profiling tools

Actual Content vs. Desired Content

Data set



SHACL Profile



SHACL Profile

Describing what ***is*** in the data set

Describing what ***should be*** in the data set

KGs Profiling tools

Users or Tools Validator

SHACL Profile: What ***is*** in the data set

	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
filter	dbo:Company	dbo:keyPerson	object								
	dbo:Company (51898)	OP dbo:keyPerson (31078)	owl:Thing	18710	29884	5263	3	1	23	2	1
	dbo:Company (51898)	OP dbo:keyPerson (31078)	foaf:Person (1179233)	7850	8333	14	1	1	15	1	1

```
Shape:Company
a sh:NodeShape;
sh:targetClass dbo:Company; #Applies to all companies.
sh:property [
        sh:name “keyPerson”;
        sh:path dbo:keyPerson; #This property shape applies to companies key person.
        sh:nodeKind sh:IRI; #The keyperson is given in IRI link
        sh:path [dbo:keyPerson
                sh:nodeShape owl:Thing;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount     “5263”^^xsd:integer;
        sh:path [sh:inversePath dbo:keyPerson;
                sh:nodeKind sh:IRI;
                sh:minCount“1” xsd:integer;
                sh:maxCount     “23” xsd:integer; ]
        ].
        sh:path [dbo:keyPerson
                sh:nodeShape foaf:Person;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount     “14”^^xsd:integer;
        sh:path [sh:inversePath dbo:keyPerson;
                sh:nodeKind sh:IRI;
                sh:minCount “1” xsd:integer;
                sh:maxCount     “15” xsd:integer; ]
        ].
```

SHACL Profile: What *is* in the data set

	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
filter	dbo:Company	dbo:keyPerson	object								
	dbo:Company (51898)	OP dbo:keyPerson (31078)	owl:Thing	18710	29884	5263	3	1	23	2	1
	dbo:Company (51898)	OP dbo:keyPerson (31078)	foaf:Person (179233)	7850	8333	14	1	1	15	1	1

```
Shape Company
a sh:NodeShape;
sh:targetClass dbo:Company; #Applies to all companies.
sh:property [
        sh:name “keyPerson”;
        sh:path dbo:keyPerson; #This property shape applies to companies key person.
        sh:nodeKind sh:IRI; #The keyperson is given in IRI link
        sh:path [dbo:keyPerson
                sh:nodeShape owl:Thing;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount     “5263”^^xsd:integer;
        sh:path [sh:inversePath dbo:keyPerson;
                sh:nodeKind sh:IRI;
                sh:minCount“1” xsd:integer;
                sh:maxCount     “23” xsd:integer; ]
        ].
        sh:path [dbo:keyPerson
                sh:nodeShape foaf:Person;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount     “14”^^xsd:integer;
        sh:path [sh:inversePath dbo:keyPerson;
                sh:nodeKind sh:IRI;
                sh:minCount “1” xsd:integer;
                sh:maxCount     “15” xsd:integer; ]
        ].
```

SHACL Profile: What ***is*** in the data set

	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
filter	dbo:Company	dbo:keyPerson	object								
	dbo:Company (51898)	OP dbo:keyPerson (31078)	owl:Thing	18710	29884	5263	3	1	23	2	1
	dbo:Company (51898)	OP dbo:keyPerson (31078)	foaf:Person (1179233)	7850	8333	14	1	1	15	1	1

```
Shape Company
a sh:NodeShape;
sh:targetClass dbo:Company; #Applies to all companies.
sh:property [
        sh:name “keyPerson”;
        sh:path dbo:keyPerson; #This property shape applies to companies key person.
        sh:nodeKind sh:IRI; #The keyperson is given in IRI link
        sh:path [dbo:keyPerson
                sh:nodeShape owl:Thing;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount “5263”^^xsd:integer;
        sh:severity sh:Warning ;
        sh:sparql [
                sh:message “Triples that might violate quality”;
                sh:prefixes dbo: ;
                sh:select “””
                SELECT ?s ?o
                WHERE {
                ?o dbo:keyPerson ?s;
                a dbo:Company .
                FILTER NOT EXISTS {
                ?s a ?type .
  }
}               “””;
        ].
```

SHACL Profile: What *is* in the data set

	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
filter	dbo:Company	dbo:keyPerson	object								
	dbo:Company (51898)	OP dbo:keyPerson (31078)	owl:Thing	18710	29884	5263	3	1	23	2	1
	dbo:Company (51898)	OP dbo:keyPerson (31078)	foaf:Person (1179233)	7850	8333	14	1	1	15	1	1

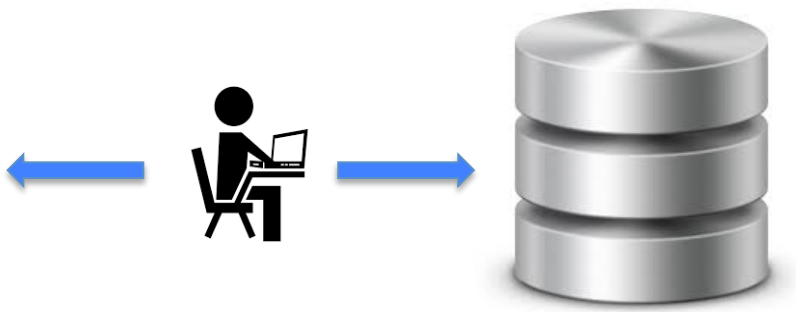
```
Shape Company
a sh:NodeShape;
sh:targetClass dbo:Company; #Applies to all companies.
sh:property [
        sh:name “keyPerson”;
        sh:path dbo:keyPerson; #This property shape applies to companies key person.
        sh:nodeKind sh:IRI; #The keyperson is given in IRI link
        sh:path [dbo:keyPerson
                sh:nodeShape owl:Thing;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount “5263”^^xsd:integer;
        sh:severity sh:Warning ;
        sh:sparql [
                sh:message “Triples that might violate quality”;
                sh:prefixes dbo: ;
                sh:select “””
                SELECT ?s ?o
                WHERE {
                ?o dbo:keyPerson ?s;
                a dbo:Company .
                FILTER NOT EXISTS {
                ?s a ?type .
 }
}               “””;
        ].
```

<dbr:Kodak dbo:keyPerson dbr:Chief_executive_officer>
<dbr:Telefonica dbo:keyPerson dbr:Chief_executive_officer>
<dbr:Allianz dbo:keyPerson dbr:Chief_executive_officer>

What ***should be*** in the data – manual setting of SHACL Constraints

	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
filter	dbo:Company	dbo:keyPerson	object								
	dbo:Company (51898)	OP dbo:keyPerson (31078)	owl:Thing	18710	29884	5263	3	1	23	2	1
	dbo:Company (51898)	OP dbo:keyPerson (31078)	foaf:Person (1179233)	7850	8333	14	1	1	15	1	1

```
Shape:Company
a sh:NodeShape;
sh:targetClass dbo:Company; #Applies to all companies.
sh:property [
        sh:name “keyPerson”;
        sh:path dbo:keyPerson; #This property shape applies to companies key person.
        sh:nodeKind sh:IRI; #The keyperson is given in IRI link
        sh:path [dbo:keyPerson
                  sh:nodeShape owl:Thing;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount “3”^^xsd:integer;
        sh:path [sh:inversePath dbo:keyPerson;
                sh:nodeKind sh:IRI;
                sh:minCount “1” xsd:integer;
                sh:maxCount “1” xsd:integer; ]
        ].
        sh:path [dbo:keyPerson
                 sh:nodeShape foaf:Person;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount “1”^^xsd:integer;
        sh:path [sh:inversePath dbo:keyPerson;
                sh:nodeKind sh:IRI;
                sh:minCount “1” xsd:integer;
                sh:maxCount “1” xsd:integer; ]
        ].
```

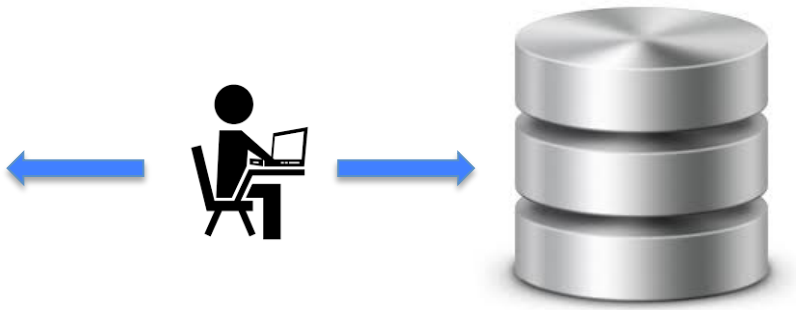


Data set

What ***should be*** in the data – heuristic generation of SHACL Constraints

	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
filter	dbo:Company	dbo:keyPerson	object								
	dbo:Company (51898)	OP dbo:keyPerson (31078)	owl:Thing	18710	29884	5263	3	1	23	2	1
	dbo:Company (51898)	OP dbo:keyPerson (31078)	foaf:Person (1179233)	7850	8333	14	1	1	15	1	1

```
Shape:Company
a sh:NodeShape;
sh:targetClass dbo:Company; #Applies to all companies.
sh:property [
        sh:name “keyPerson”;
        sh:path dbo:keyPerson; #This property shape applies to companies key person.
        sh:nodeKind sh:IRI; #The keyperson is given in IRI link
        sh:path [dbo:keyPerson
                  sh:nodeShape owl:Thing;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount “6”^^xsd:integer;
        sh:path [sh:inversePath dbo:keyPerson;
                sh:nodeKind sh:IRI;
                sh:minCount “1” xsd:integer;
                sh:maxCount “4” xsd:integer; ]
        ].
        sh:path [dbo:keyPerson
                 sh:nodeShape foaf:Person;
        sh:minCount “1”^^xsd:integer; #keyPerson is a required property.
        sh:maxCount “1”^^xsd:integer;
        sh:path [sh:inversePath dbo:keyPerson;
                sh:nodeKind sh:IRI;
                sh:minCount “1” xsd:integer;
                sh:maxCount “1” xsd:integer; ]
        ].
```

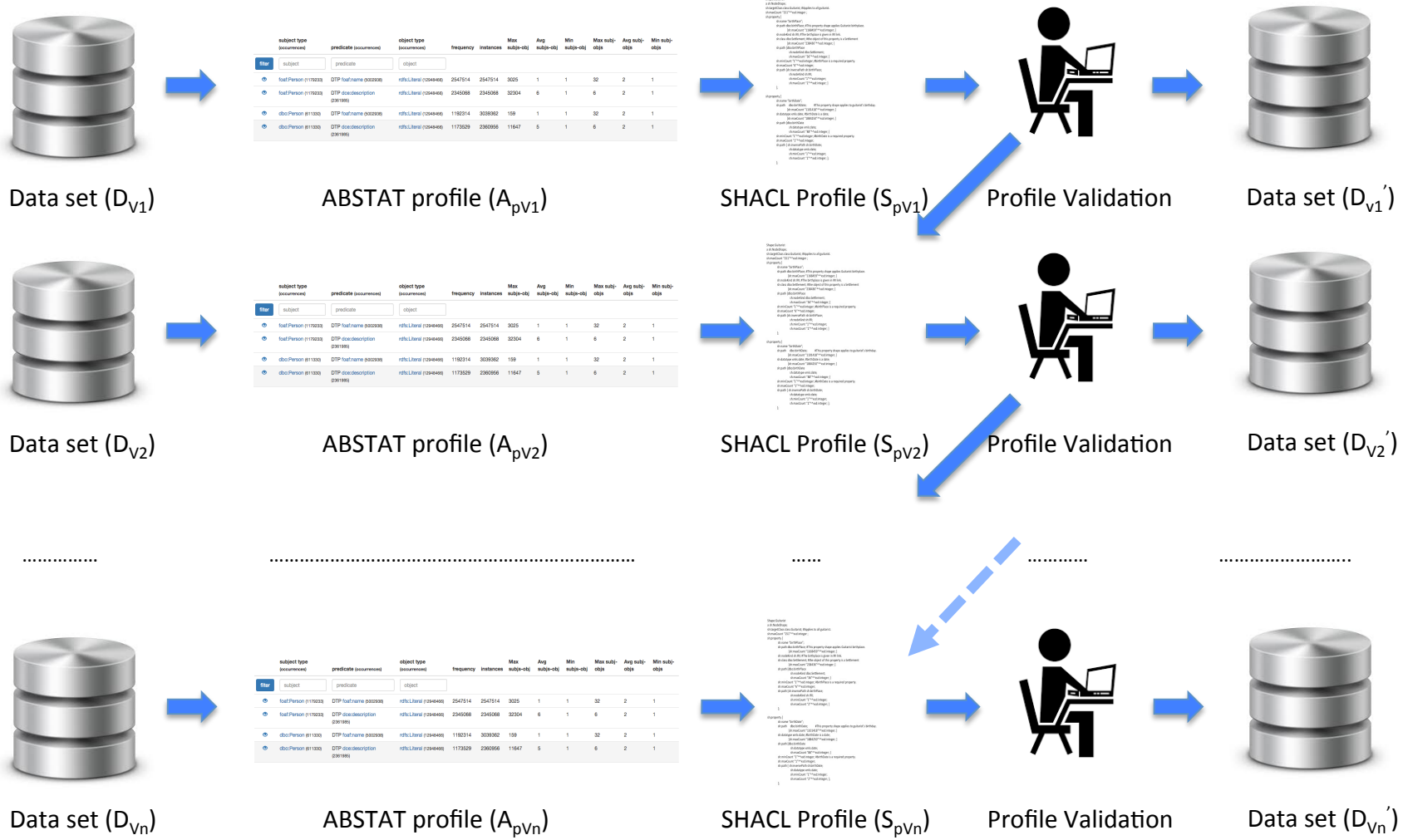


Data set

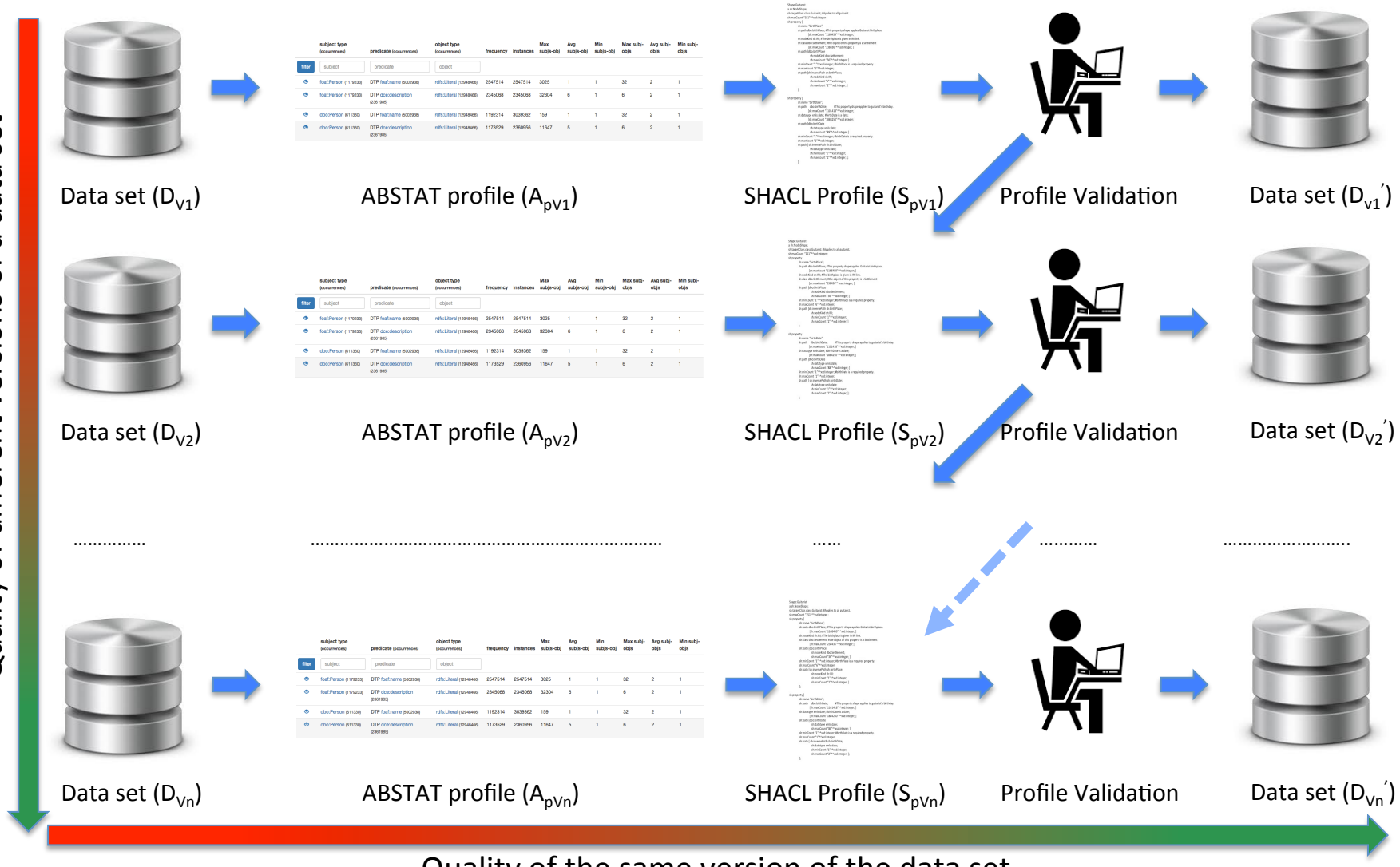
2*Avg subjs-obj

n*Avg subjs-obj

SHACL Generation and Validation Methodology



SHACL Generation and Validation Methodology



Conclusions and Future Work

Take home message:

- A methodology for assessing the quality of the data set and its versions
- Automatic generation of SHACL Profile with heuristic setting of cardinality constraints

"Towards":

- Better heuristics (feedback at this workshop will be appreciated ☺)
- Integration with SHACL validation tool for full methodology implementation
- Run the experiment in large scale

Università degli Studi di Milano – Bicocca
Dipartimento di Informatica Sistemistica e Comunicazione

THANK YOU FOR YOUR ATTENTION!

INSID&S Lab
Interaction and Semantics
for Innovation with Data & Services

@InsideLaBicocca
@blerinaspahiu