On Learnability of Constraints from RDF Data

Emir Muñoz

Fujitsu Ireland Ltd.
Insight Centre for Data Analytics, NUI Galway

ESWC 2016 PhD Symposium

Motivation (1/6)

Resource Description Framework (RDF) is …

Structured data

Dynamic data

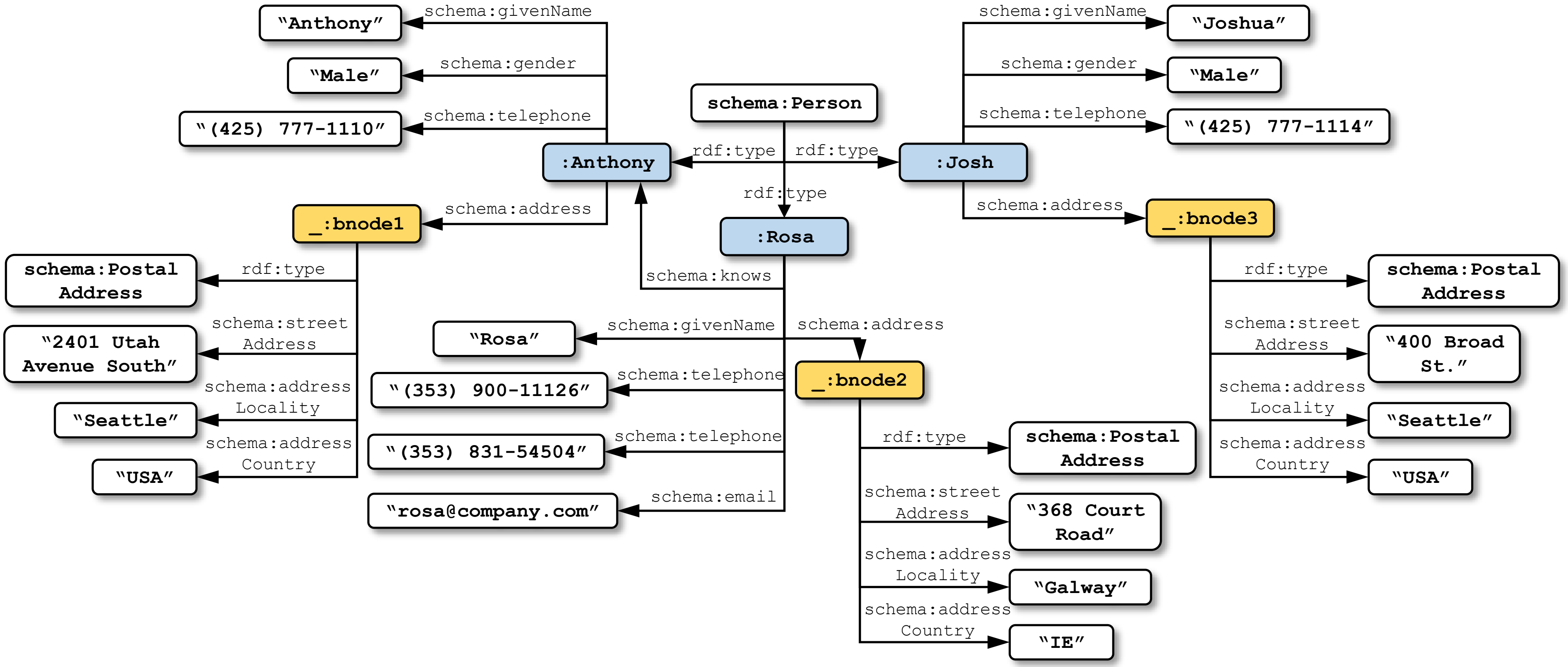
Schema-less data

Good for the Web (data integration, transfer, etc.)

Bad for users (reusability, trust, understanding, etc.)

Challenges arise due to the **Open World Assumption** (OWA) and **non-Unique Name Assumption** (nUNA) in OWL/RDF

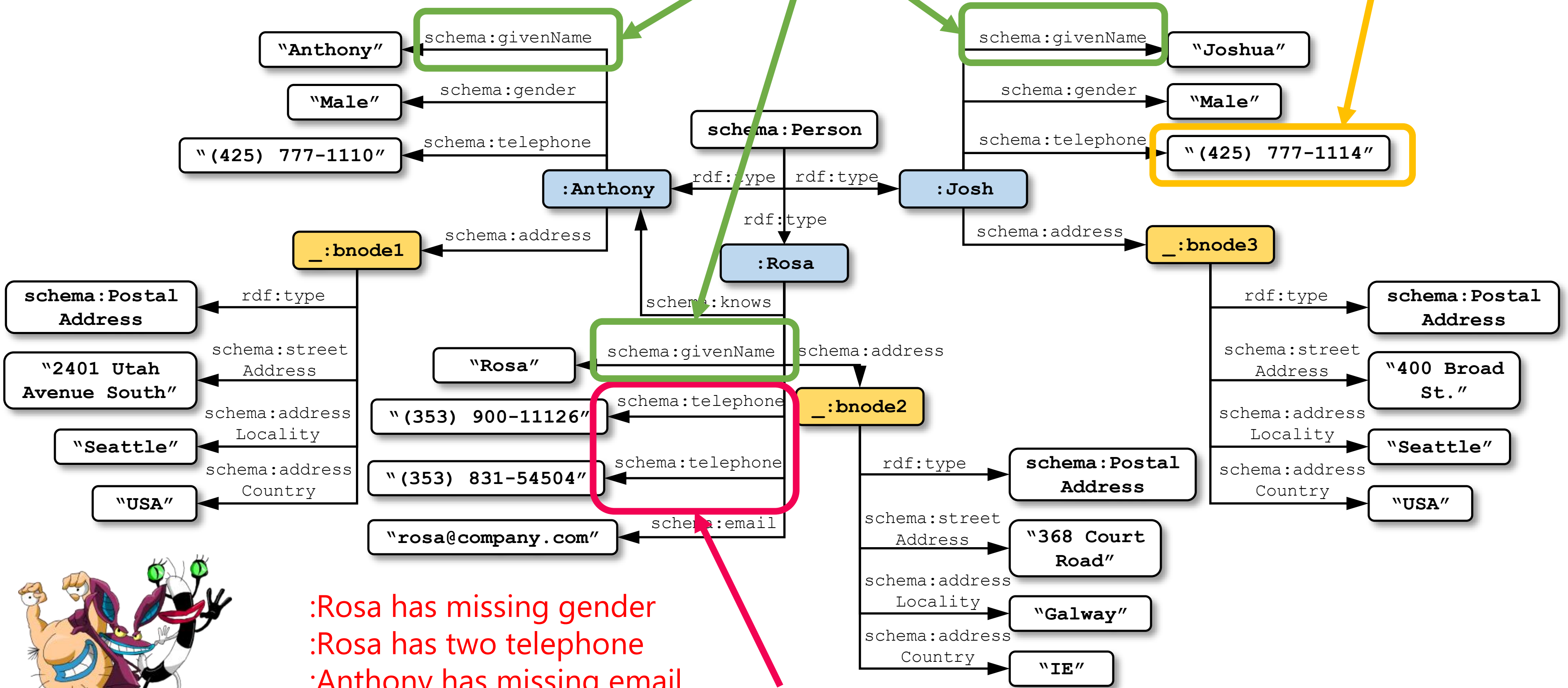
Motivation (2/6)



Motivation (2/6)

Exactly one value (key)

It follows a syntactic pattern



:Rosa has missing gender
:Rosa has two telephone
:Anthony has missing email
:Josh has missing email

Cardinality (max) 2 (not given by schema.org)

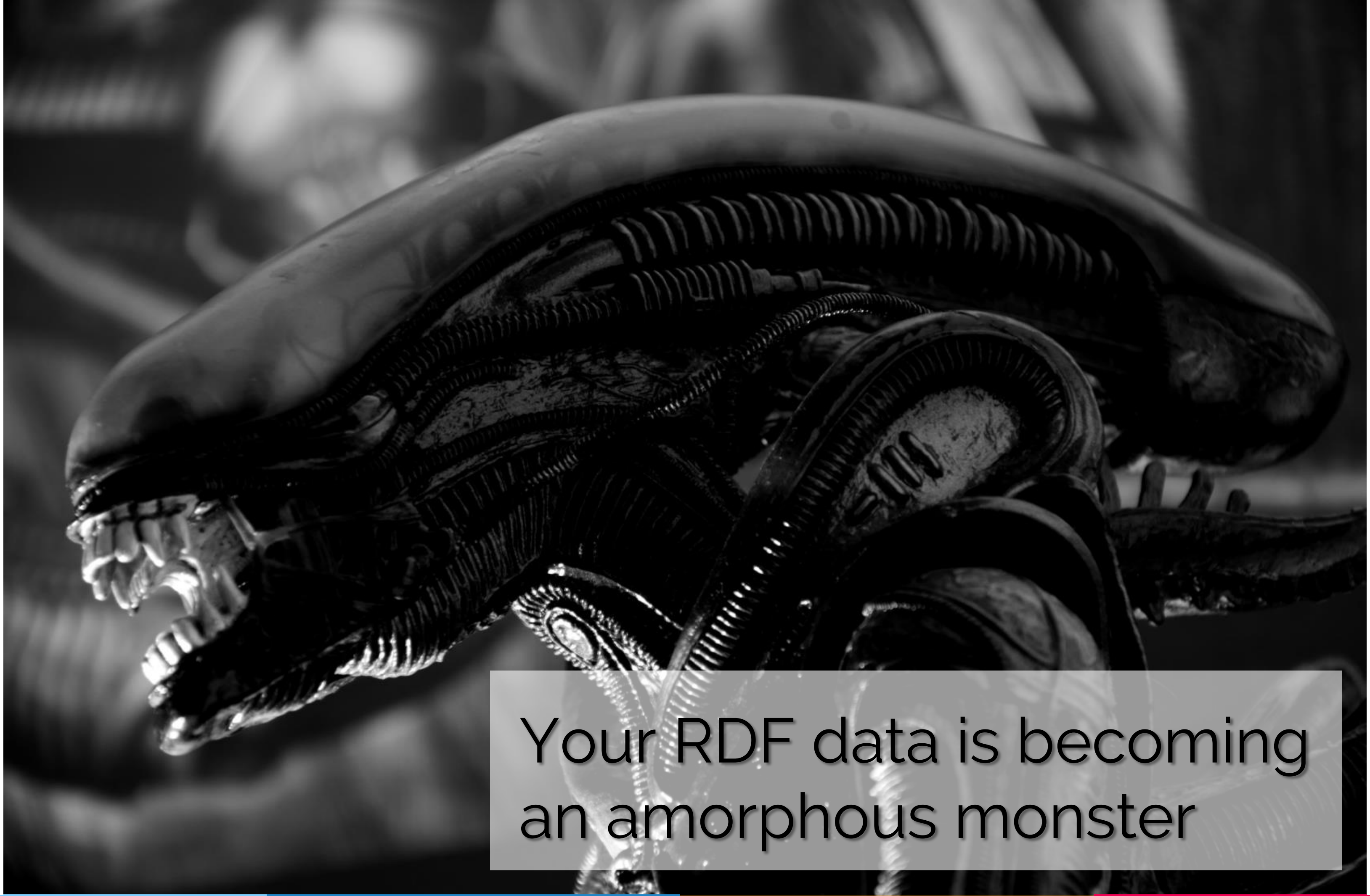
Motivation (3/6)

- Such restrictions are required while querying RDF
- Even when ontologies or vocabularies are present!
- Without knowledge about the instance data
  - user cannot be sure which predicates are present (e.g., `schema:email`)
  - or which of them are multi-valued (e.g., `schema:telephone`)

```
SELECT ?person ?givenName (GROUP_CONCAT(?email; separator=", ") AS ?email)
WHERE {
     ?person rdf:type schema:Person .
     OPTIONAL { ?person schema:givenName ?givenName }
     OPTIONAL { ?person schema:email ?email }
} GROUP BY ?person ?givenName
```

Similar example was used as motivation in [1]

[1] G. Lausen, M. Meier, and M. Schmidt. *SPARQLing constraints for RDF*. EDBT 2008.


Your RDF data is becoming
an amorphous monster

If RDF is schema less… how can I know the structure of my data?

RDF KG = {RDF triples} that “follow” an implicit **schema structure**

We could then learn the characteristics of RDF data under a **Closed World Assumption** (CWA) with UNA

Motivation (6/6)

- Constraints can help to represent characteristics that data naturally exhibits
  - Every person contains exactly one value for the `schema:givenName` and `schema:address` properties
  - The combines properties `schema:givenName` and `schema:address` uniquely identify each person in the data
  - Each person is connected to at least one value for the `schema:telephone` property and at most two values
  - All values of the property `schema:telephone` follow the same '(NUMBER NUMBER-NUMBER)' syntactic pattern
  - Entities with a `schema:givenName` and `schema:address` must be instances of the class `schema:Person`

State-of-the-art (1/2)

- Constraints are limitations incorporated on the data that are supposed to be satisfied all the time

  Types: Integrity, Cardinality, Type, Domain/Range, etc.
- Very common in relational databases
- First introduced to RDF by Lausen et at. [1] in 2008

  Goal: Convert RDB to RDF without losing semantic information
- OWL 2 allows the definition of some constraints: `owl:hasKey`, `owl:minCardinality/maxCardinality/exactCardinality`
- However, ontologies constrain the domain *not* the data

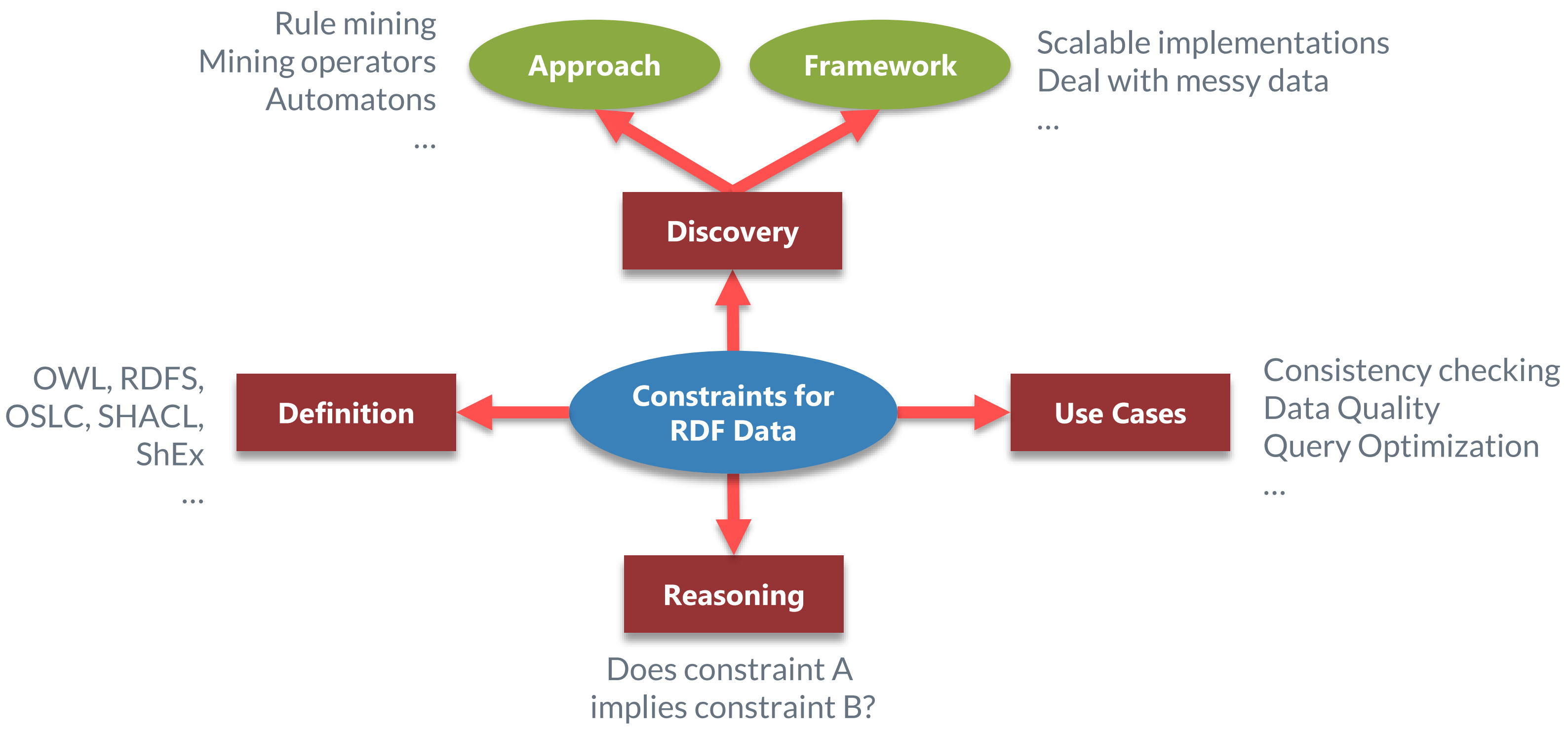
[1] G. Lausen, M. Meier, and M. Schmidt. *SPARQLing constraints for RDF*. EDBT 2008.

State-of-the-art (2/2)

- Brand new Constraint Languages for RDF: ShEx[2], RDD[3], SHACL[4], SPIN[5], OSLC[6]
- Designed for validation against a user-defined "shape"
- Main drawbacks:
  - Users should define the constraints
  - Low expressivity of defined constraints in general
  - Not widely adopted yet

[2] https://www.w3.org/2013/ShEx/Primer
[3] P. M. Fischer, G. Lausen, A. Schatzle, and M. Schmidt. *RDF Constraint Checking*. EDBT/ICDT Workshops 2015.
[4] https://www.w3.org/TR/shacl/
[5] http://spinrdf.org/
[6] https://www.w3.org/Submission/2014/SUBM-shapes-20140211/

Problem Statement and Contributions (1/2)



Problem Statement and Contributions (2/2)

RQ1: Can we define expressive and novel constraints for RDF data?

Definition ← Constraints for RDF Data

RQ2: Can constraints be automatically extracted under a non-CBD assumption?

Discovery ← Constraints for RDF Data

RQ3: What is the impact of constraints in the assessment of RDF data quality?

Use Cases ← Constraints for RDF Data

CBD - Concise Bounded Description (https://www.w3.org/Submission/CBD/)

Methodology (1/3)

Definition of constraints for RDF

- Consider Blank Nodes
- Increase expressivity with SPARQL Property Paths[7]

  `schema:address/schema:streetAddress`
- Notion of soft or probability constraints to avoid data loss

[7] https://www.w3.org/TR/sparql11-property-paths/

Methodology (2/3)

Discovery of constraints for RDF

- Approaches to discover some of these constraints
- How to deal with different modellings (e.g., CBD*)?
- Translation of XML and RDB approaches
- Scalability to support large-scale RDF datasets

(*) Non standard RDF summarization

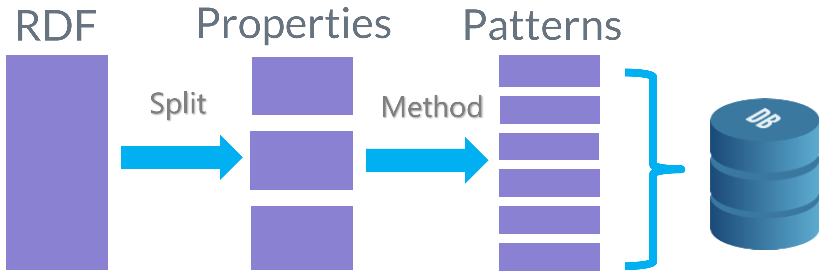
Methodology (3/3)

Constraints and Data Quality

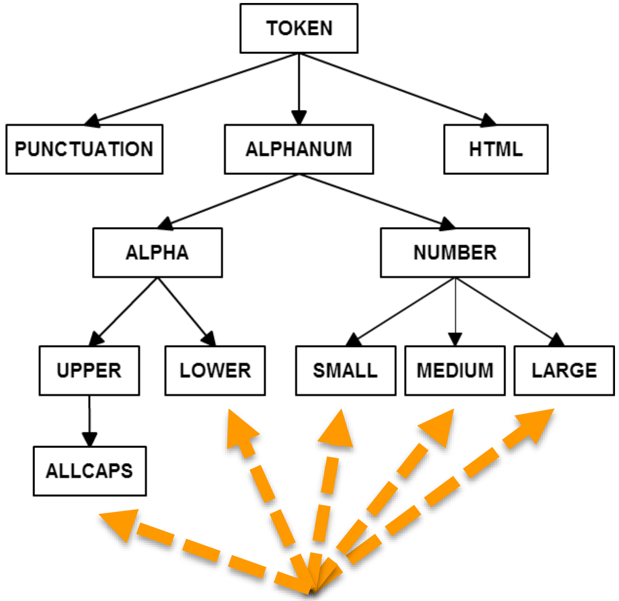
- Constraints could be related with several data quality dimensions
- Practical study on the benefits of constraints

Preliminary Results (1/2)

- Syntactic pattern constraints
- Limited to literal values



Lerman et al.[8]



More specific categories

[8] K. Lerman, S. Minton, and C.A. Knoblock. *Wrapper Maintenance: A Machine Learning Approach*. JAIR 2003.

Preliminary Results (2/2)

- 500k patterns in our database coming from DBpedia
- Different use cases:
  - Search for properties
  - Validation of values
  - Information extraction based on patterns

```
vcard:email  mailto : ALPHA PUNCTUATION ALL_LOWERCASE . ALL_LOWERCASE        0.82
vcard:email  mailto : ALPHA PUNCTUATION ALL_LOWERCASE . com                  0.69
vcard:email  mailto : ALPHA @ ALPHANUMERIC . ALL_LOWERCASE                   0.54
vcard:email  mailto : ALPHA @ ALPHANUMERIC . com                             0.46
vcard:email  mailto : ALL_UPPERCASE ****@ ALL_LOWERCASE . ALL_LOWERCASE      0.36
```

Evaluation Plan (1/3)

Definition of constraints for RDF

- Comparison of the expressivity of current definitions against the new ones that involve SPARQL Property Paths
- Compare against semantically similar definitions in XML and RDBs

Evaluation Plan (2/3)

Discovery of constraints for RDF

- For key constraints compare against ROCKER[9]
- Build manually annotated gold-standard
    - A source could be Web Data Commons[10]
    - RDF benchmarks
- Test scalability in different size datasets

[9] T. Soru, E. Marx, and A.-C. Ngonga Ngomo. *ROCKER -- A Refinement Operator for Key Discovery*. WWW 2015.
[10] http://webdatacommons.org/

Evaluation Plan (3/3)

Constraints and Data Quality

- Carry out the validation of our constraints against the source dataset (division in train/set set)
  - Make use of ShEx or RDD implementations
- User study to determine usefulness of extracted constraints. *Does a constraint match any business rule?*

Summary

- RDF constraints are limited by their mapping from RDBs
- They do not consider complex values or graph nature of RDF

  e.g., Keys are defined as a set of properties
- We aim to unlock further applications in data cleaning, integration, modeling, processing, and retrieval akin to constraints in RDBs

Thanks!

Any questions?

Emir Muñoz
emir@emunoz.org

APPENDICES

▷ RDD vs Shape Expressions[3]

RDD

```
OWA CLASS foaf:Person {
   KEY rdfs:label : LITERAL
   MAX(2) foaf:mbox : LITERAL
   TOTAL foaf:age : LITERAL(xsd:int)
   RANGE(foaf:Person) foaf:knows : IRI
}
```

- More focus on verification
- Inspired by relational constraints
- Validation of typed datasets
- Meaning: Are there instances of type *person* that do not adhere to the schema?

Shape Expressions (ShEx)

```
<Person> {
   KEY rdfs:label xsd:string ,
   MAX foaf:mbox xsd:string{0,2} ,
   TOTAL foaf:age xsd:int ,
   RANGE foaf:knows @<Person>*
}
```

- More focus on type inference
- Inspired by XML RelaxNG
- Meaning: Which instances have the shape of a *person*?

[3] P. M. Fischer, G. Lausen, A. Schatzle, and M. Schmidt. *RDF Constraint Checking*. EDBT/ICDT Workshops 2015.

Concise Bounded Description (CBD)

- Given a particular node (the starting node) in a particular RDF graph (the source graph), a subgraph of that particular graph, taken to comprise a concise bounded description of the resource denoted by the starting node, can be identified as follows:

1. Include in the subgraph all statements in the source graph where the subject of the statement is the starting node;
2. Recursively, for all statements identified in the subgraph thus far having a blank node object, include in the subgraph all statements in the source graph where the subject of the statement is the blank node in question and which are not already included in the subgraph.
3. Recursively, for all statements included in the subgraph thus far, for all reifications of each statement in the source graph, include the concise bounded description beginning from the rdf:Statement node of each reification.

- This results in a subgraph where the object nodes are either URI references, literals, or blank nodes not serving as the subject of any statement in the graph.

CBD Application Issues

- Representations versus Descriptions
- Determination of the Source Graph
- Query and Application Programming Interfaces
- Managing magnitude
    - Limit the (over)use of Blank Nodes
    - Limiting Path Length
    - Limiting Total Number of Statements
    - Excluding or Limiting Reifications