KR2RML: An Alternative Interpretation of R2RML for Heterogeneous Sources

Jason Slepicka
Chengye Yin
Pedro Szekely
Craig Knoblock
What’s the problem?

- Consuming Linked Data requires RDF
- Consuming other formats requires many languages for querying, transforming, and mapping to RDF

<table>
<thead>
<tr>
<th>Source Format</th>
<th>Query Language</th>
<th>Transformation Language</th>
<th>Mapping Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDBMS</td>
<td>SQL</td>
<td>SQL</td>
<td>R2RML, D2R, RML</td>
</tr>
<tr>
<td>XML</td>
<td>XPath</td>
<td>XSLT</td>
<td>XSLT, RML, XR2RML</td>
</tr>
<tr>
<td>JSON</td>
<td>jQuery</td>
<td>JQ</td>
<td>RML, XR2RML</td>
</tr>
<tr>
<td>CSV</td>
<td>sed/awk</td>
<td>sed/awk</td>
<td>RML, XR2RML</td>
</tr>
<tr>
<td>Avro</td>
<td>HiveQL, Pig Latin</td>
<td>HiveQL, Pig Latin</td>
<td>?</td>
</tr>
<tr>
<td>Thrift</td>
<td>Hive SerDe, Pig Latin</td>
<td>HiveQL, Pig Latin</td>
<td>?</td>
</tr>
</tbody>
</table>
What would a good solution support?

• Hierarchical Input and Output Formats
• Forward Compatibility For New Formats
• Reusable Transformations
• Scalability to billions of triples
How does KR2RML (Karma R2RML) achieve these goals?
### Nested Relational Model

<table>
<thead>
<tr>
<th>companyName</th>
<th>tags</th>
<th>employees</th>
<th>locationTable</th>
</tr>
</thead>
</table>
| Information Sciences Institute | artificial intelligence, nlp, semantic web | Knoblock, Craig (Director, Research Professor) | locationAddress: 4676 Admiralty Way Suite 1001, Marina Del Rey, CA 90292  
|                 |               | Slepicka, Jason (Graduate Student, Research Assistant) | values: ISI - West, ISI - East |
| Institute for Creative Technologies | computer graphics, virtual reality | Debevec, Paul (Research Professor, Associate Director) | locationName: 12015 Waterfront Drive, Playa Vista, CA, 90094  
|                 |               | Swartout, William (Chief Technology Officer, Research) | values: ICT - Headquarters |
Transformations

• Structural
  – Split, Glue, Fold, Unfold,

• Value
  – Python User Defined Functions and Aggregations

• Filters
Transformation Example: Split

<table>
<thead>
<tr>
<th>employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
</tr>
<tr>
<td>title</td>
</tr>
<tr>
<td>Roles</td>
</tr>
<tr>
<td>Values</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knoblock, Craig</th>
<th>Director, Research Professor</th>
<th>Director</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Research Professor</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Slepicka, Jason</th>
<th>Graduate Student, Research Assistant</th>
<th>Graduate Student</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Research Assistant</td>
</tr>
</tbody>
</table>
### Transformation Examples: Glue

<table>
<thead>
<tr>
<th>locationAddress</th>
<th>locationName</th>
<th>Glue_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>4676 Admiralty Way Suite 1001, Marina Del Rey, CA</td>
<td>ISI - West</td>
<td>ISI - West</td>
</tr>
<tr>
<td>3811 North Fairfax Drive Suite 200, Arlington, VA 22203</td>
<td>ISI - East</td>
<td>4676 Admiralty Way Suite 1001, Marina Del Rey, CA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3811 North Fairfax Drive Suite 200, Arlington, VA 22203</td>
</tr>
</tbody>
</table>
Transformation Examples: Python

```python
1. return getValue("companyURI") + "/employee/" + "/".join(getValue("name").replace(' ','').split(','))
```

On Error:

Use JSON Output:

View Errors Preview results for top 5 rows

corporation/InformationSciencesInstitute/employee/Knoblock/Craig
corporation/InformationSciencesInstitute/employee/Siepicka/Jason
### Transformation Examples: Python

<table>
<thead>
<tr>
<th>Location</th>
<th>Address</th>
<th>ZIP Code</th>
<th>State</th>
<th>City</th>
<th>Street Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISI - West</td>
<td>4676 Admiralty Way Suite 1001,</td>
<td>90292</td>
<td>CA</td>
<td>Marina Del Rey</td>
<td>4676 Admiralty Way Suite 1001</td>
</tr>
<tr>
<td>ISI - East</td>
<td>3811 North Fairfax Drive Suite 200,</td>
<td>22203</td>
<td>VA</td>
<td>Arlington</td>
<td>3811 North Fairfax Drive Suite 200</td>
</tr>
</tbody>
</table>
R2RML Applied to Relational Data Model
R2RML Applied to Relational Data Model
KR2RML applied to Nested Relational Model
KR2RML applied to Nested Relational Model

_:TriplesMap_1

_:SubjectMap_1

rr:class

schema:Person

_:PredicateObjectMap_1

rr:predicate

schema:name

_:ObjectMap_1

["employees","name"]

rr:column
KR2RML Processing

RDF Generation Triples Map Processing Order

_:TriplesMap_4 (PostalAddress1)  _:TriplesMap_3 (Place1)  _:TriplesMap_2 (Person1)*  _:TriplesMap_1 (Organization1)
KR2RML Processing: ObjectMap
KR2RML Processing: RefObjectMap
KR2RML JSON-LD Output

```json
{
"@context": "http://ex.com/contexts/iswc2015_json-context.json",
"location": [
{"address": {
"streetAddress": "4676 Admiralty Way Suite 1001",
"addressLocality": "Marina Del Rey",
"postalCode": "90292",
"addressRegion": "CA",
"a": "PostalAddress"},
"name": "ISI - West",
"a": "Place",
"uri": "isi-location:ISI-West"},
...
],
"name": "Information Sciences Institute",
"a": "Organization",
"employee": [
{"name": "Knoblock, Craig",
"a": "Person",
"uri": "isi-employee:Knoblock/Craig",
"jobTitle": ["Research Professor", "Director"],
"worksFor": "isi:company/InformationSciencesInstitute"},
...
],
"uri": "isi:company/InformationSciencesInstitute"
}
Scalability

- Disallow joins because they’re too complicated for KR2RML to come up for every big data use case
- Embedded in MapReduce and Storm
- To generate our human trafficking knowledge graph of 4 billion triples, it takes 20 machines 10 hours over 50 million documents from dozens of sources.
- That’s ~6,000 triples per second per machine!
Conclusions

• KR2RML does not require modifications to the language to support new hierarchical formats
• KR2RML mappings can be reused across source formats without modification.
• A KR2RML processor can clean and transform data in a reusable way across sources
• A KR2RML processor can materialize RDF from heterogeneous sources in streaming or batch on the order of billions of triples efficiently.
Questions?