USC INFORMATION SCIENCES INSTITUTE

KR2RML: An Alternative Interpretation of R2RML for Heterogeneous Sources

Jason Slepicka
Chengye Yin
Pedro Szekely
Craig Knoblock

What's the problem?

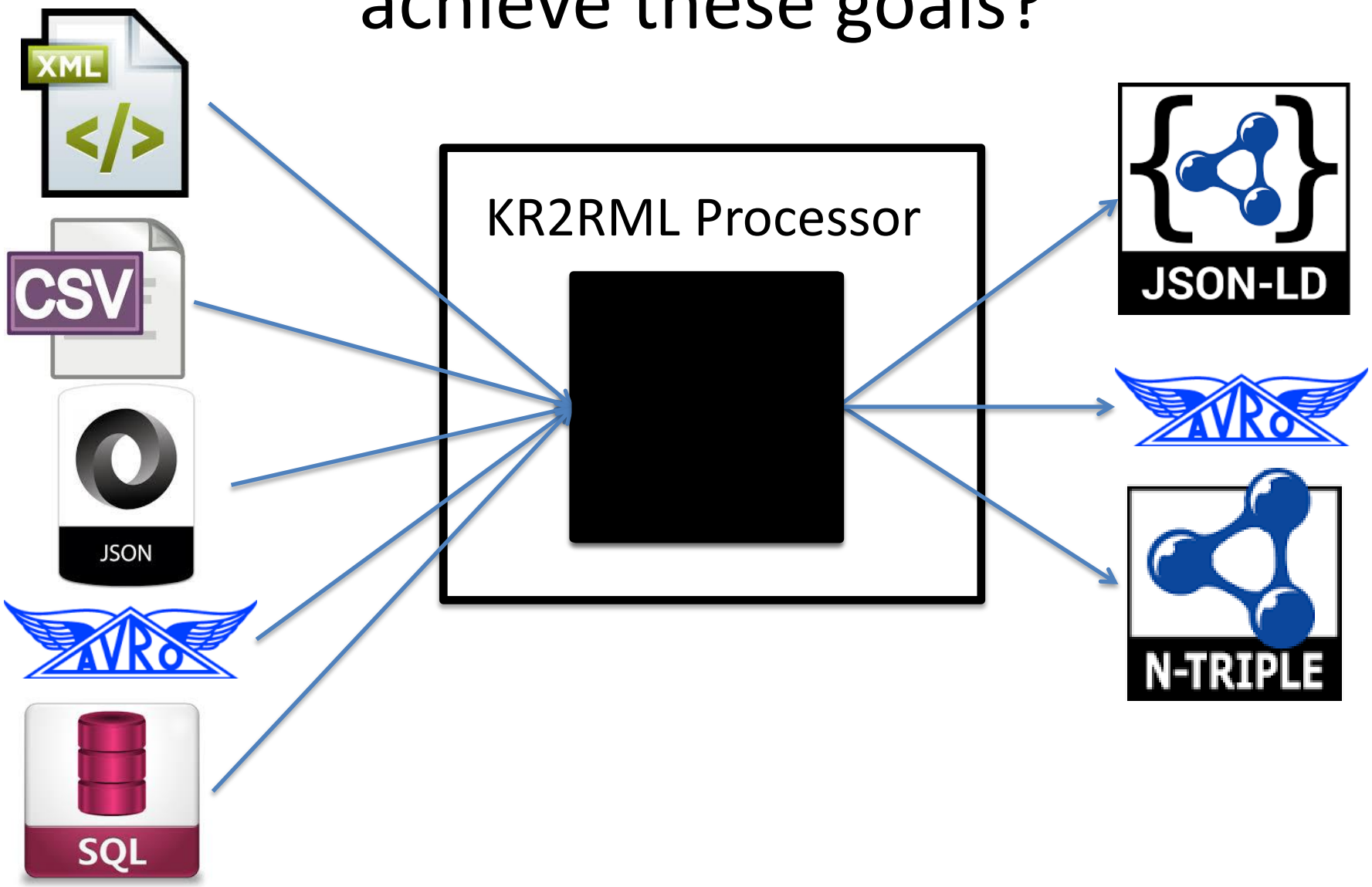
- Consuming Linked Data requires RDF
- Consuming other formats requires many languages for querying, transforming, and mapping to RDF

Source Format	Query Language	Transformation Language	Mapping Language
RDBMS	SQL	SQL	R2RML, D2R, RML
XML	XPath	XSLT	XSLT, RML, XR2RML
JSON	jQuery	JQ	RML, XR2RML
CSV	sed/awk	sed/awk	RML, XR2RML
Avro	HiveQL, Pig Latin	HiveQL, Pig Latin	?
Thrift	Hive SerDe, Pig Latin	HiveQL, Pig Latin	?

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

companyName	tags	employees		locationTable	
	values	name	title	locationAddress	locationName
				values	values
Information Sciences Institute	artificial intelligence nlp semantic web	Knoblock, Craig	Director, Research Professor	4676 Admiralty Way Suite 1001, Marina Del Rey, CA 90292 3811 North Fairfax Drive Suite 200, Arlington, VA	ISI - West ISI - East
		Slepicka, Jason	Graduate Student, Research Assistant		
Institute for Creative Technologies	computer graphics virtual reality	Debevec, Paul	Research Professor, Associate Director	12015 Waterfront Drive, Playa Vista, CA, 90094	ICT - Headquarters
		Swartout, William	Chief Technology Officer, Research		

Transformations

- Structural
  - Split, Glue, Fold, Unfold,
- Value
  - Python User Defined Functions and Aggregations
- Filters

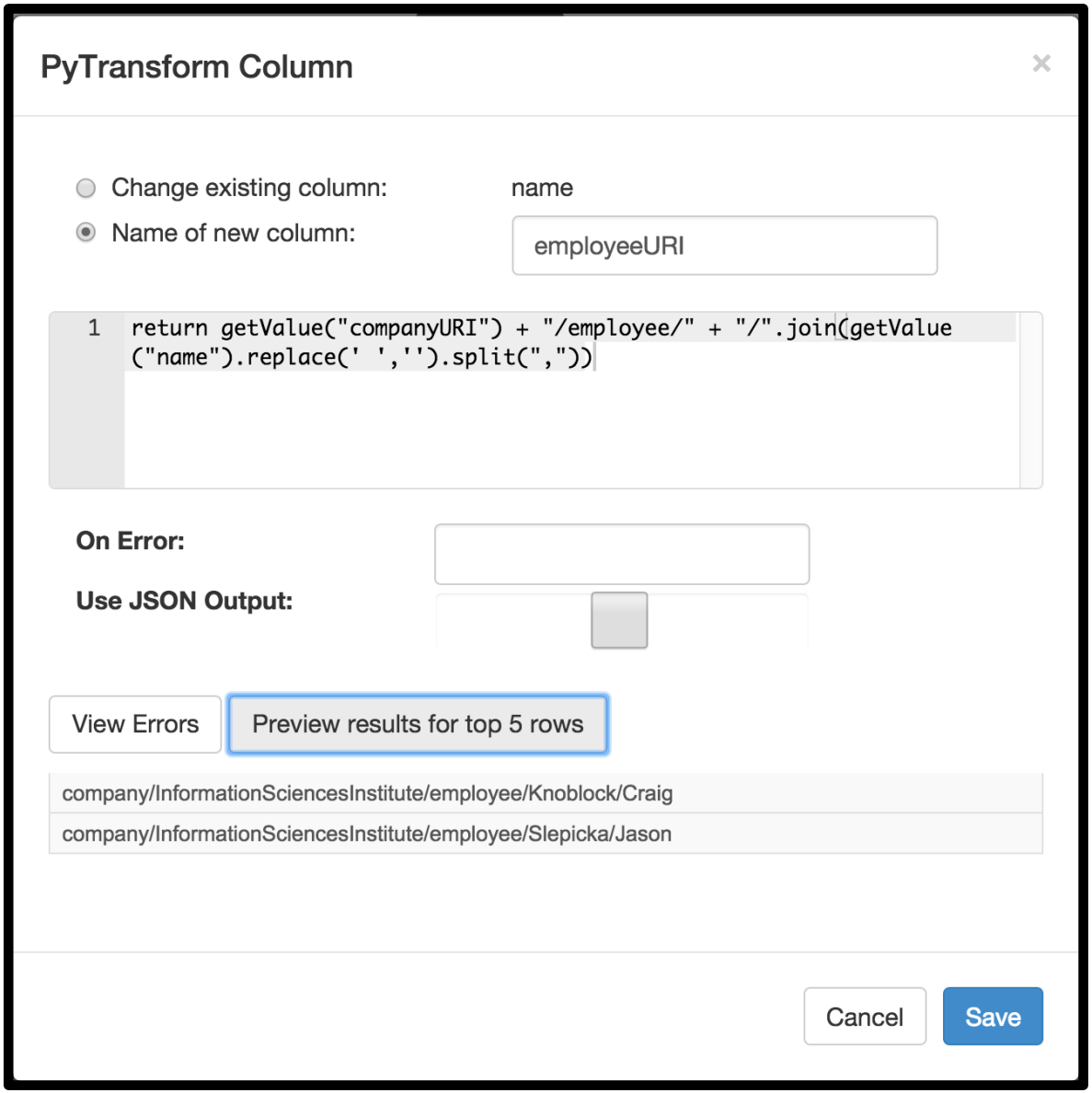
Transformation Example: Split

employees		
name	title	Roles Values
Knoblock, Craig	Director, Research Professor	Director Research Professor
Slepicka, Jason	Graduate Student, Research Assistant	Graduate Student Research Assistant

Transformation Examples: Glue

locationTable			
locationAddress	locationName	Glue_1	
values	values	values	values_1
4676 Admiralty Way Suite 1001, Marina Del Rey, CA	ISI - West	ISI - West	4676 Admiralty Way Suite 1001, Marina Del Rey, CA
3811 North Fairfax Drive Suite 200, Arlington, VA 22203	ISI - East	ISI - East	3811 North Fairfax Drive Suite 200, Arlington, VA 22203

Transformation Examples: Python



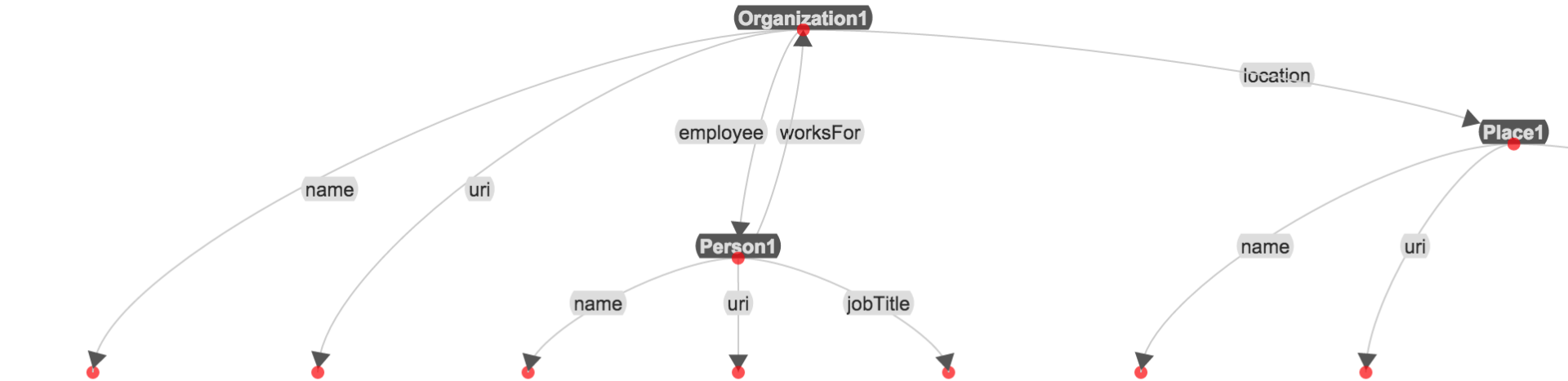
Transformation Examples: Python

locationTable

Glue_1

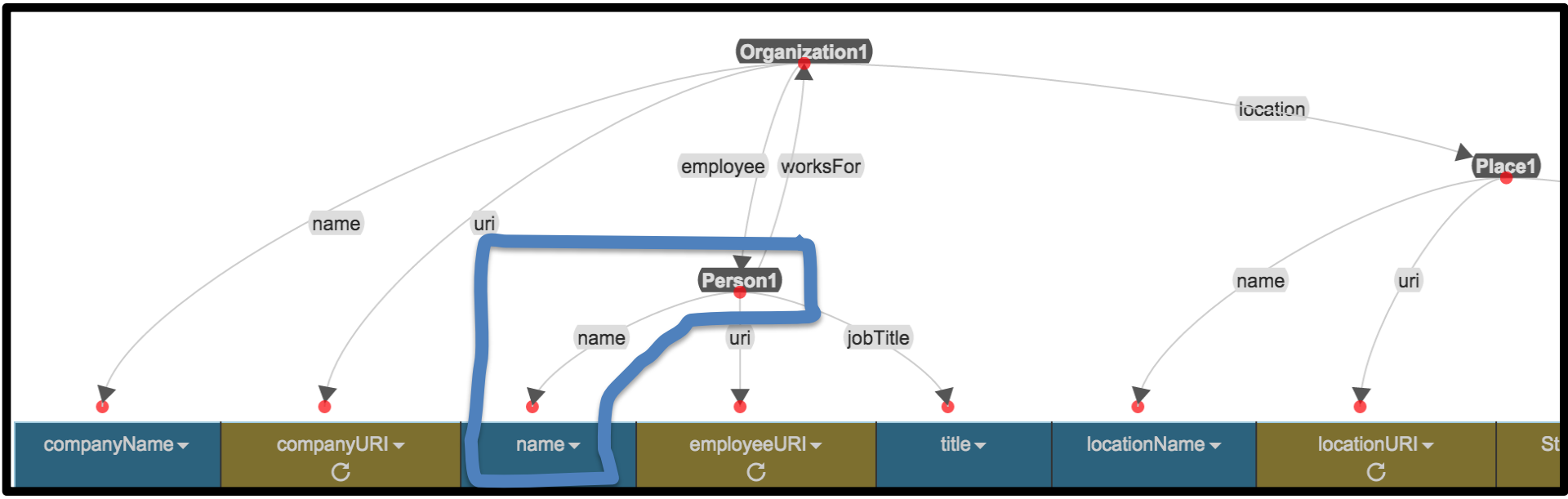
values	values_1	ZIP Code	State	City	Street Address
ISI - West	4676 Admiralty Way Suite 1001,	90292	CA	Marina Del Rey	4676 Admiralty Way Suite 1001
ISI - East	3811 North Fairfax Drive Suite 200,	22203	VA	Arlington	3811 North Fairfax Drive Suite 200

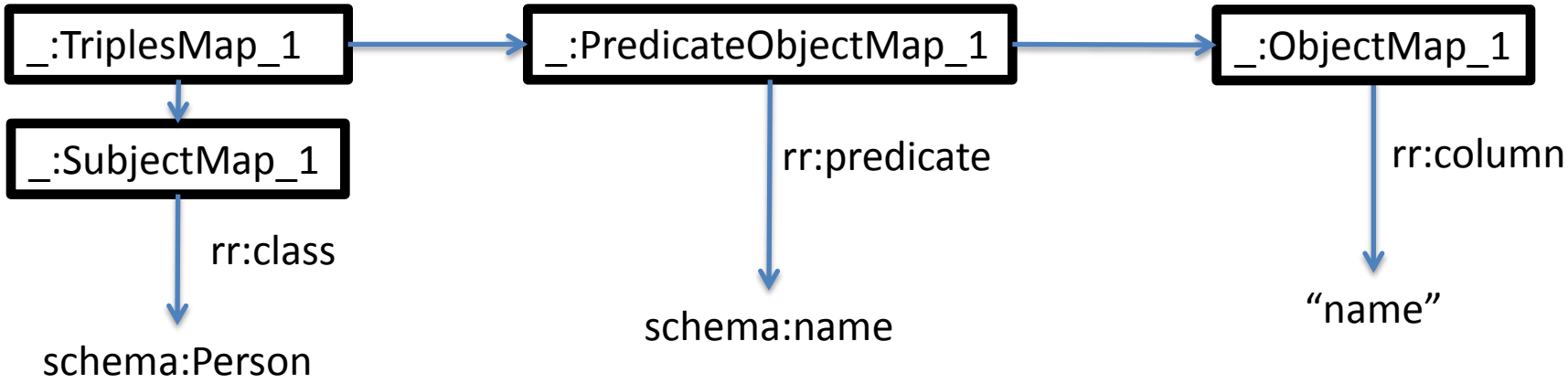
R2RML Applied to Relational Data Model



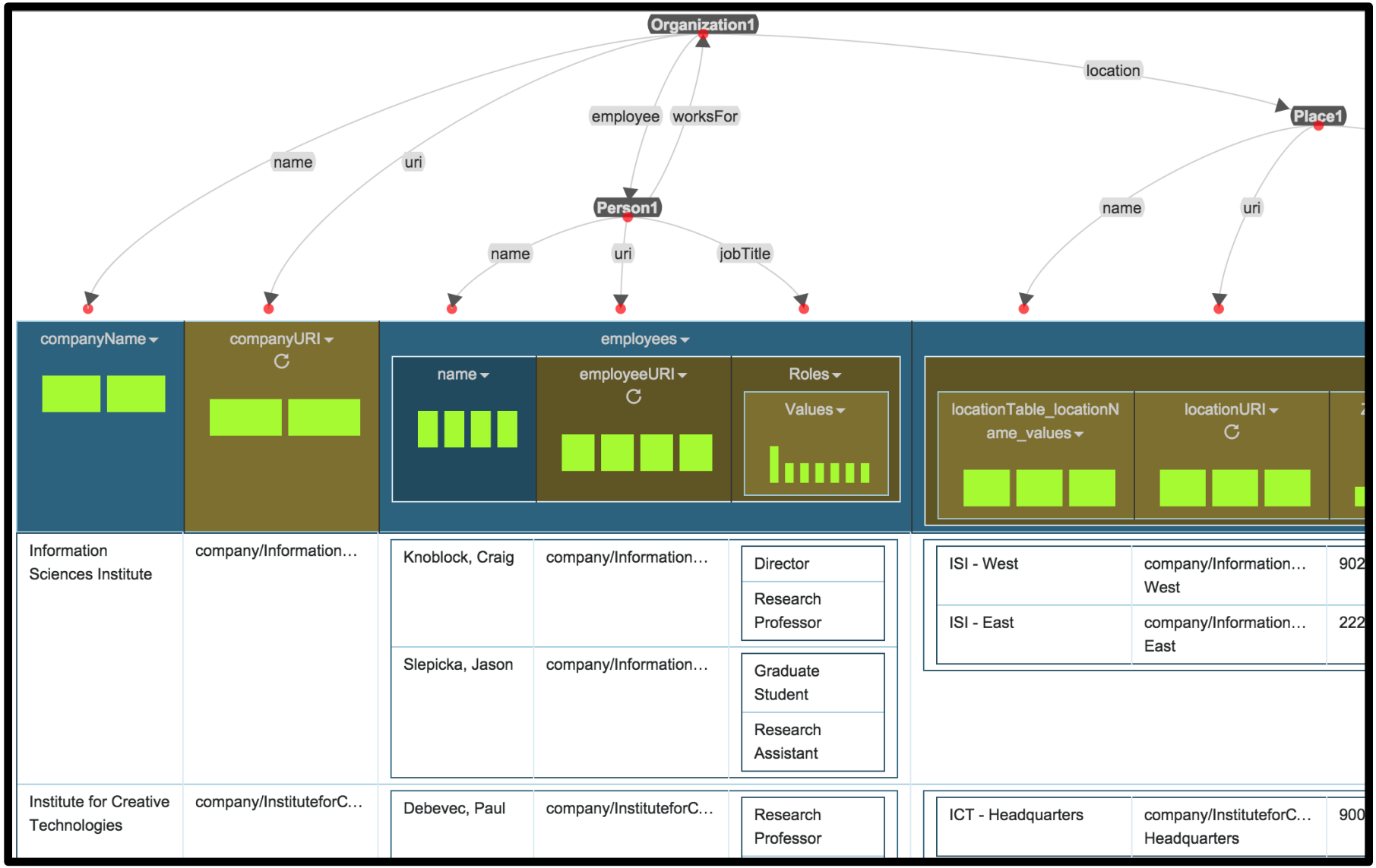
companyName	companyURI	name	employeeURI	title	locationName	locationURI	St
Information Sciences Institute	company/Information…	Knoblock, Craig	company/Information…	Director	ISI - West	company/Information…	4676 Suite
Institute for Creative Technologies	company/InstituteforC…	Debevec, Paul	company/InstituteforC…	Research Professor	ICT - Headquarters	company/InstituteforC…	12015

R2RML Applied to Relational Data Model

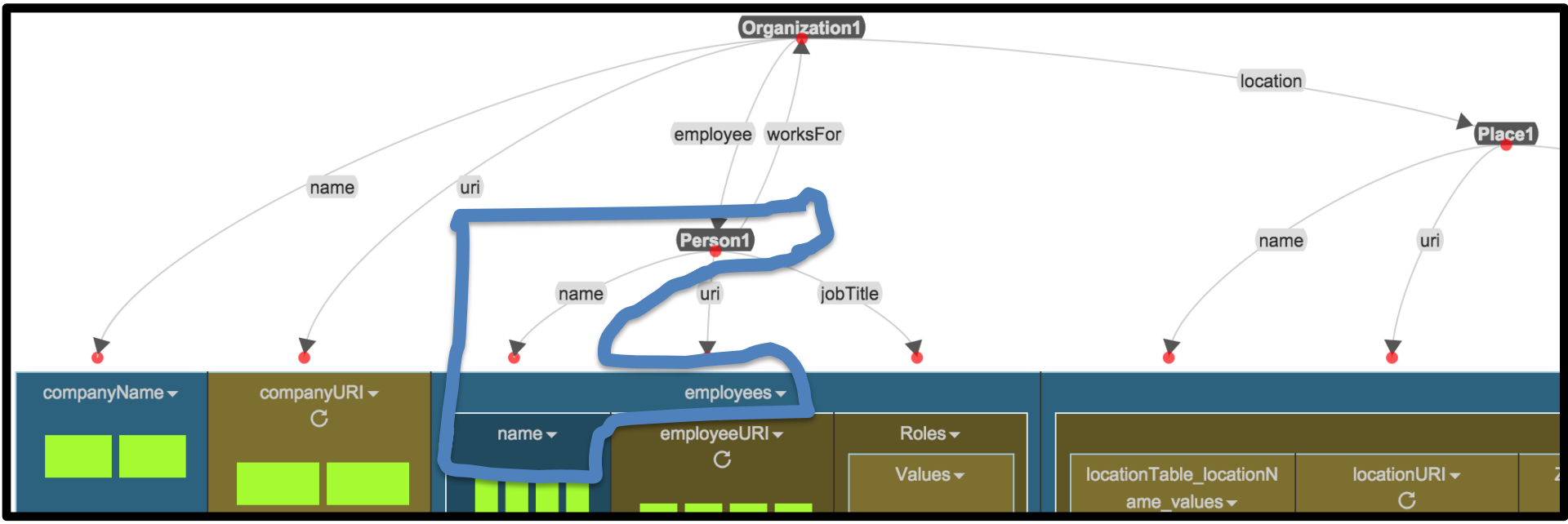


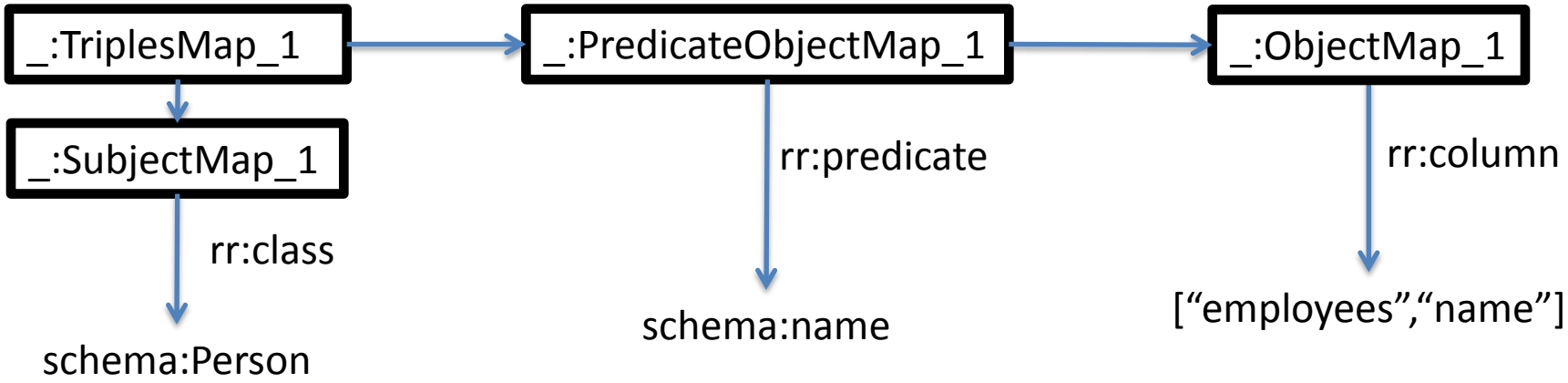


KR2RML applied to Nested Relational Model

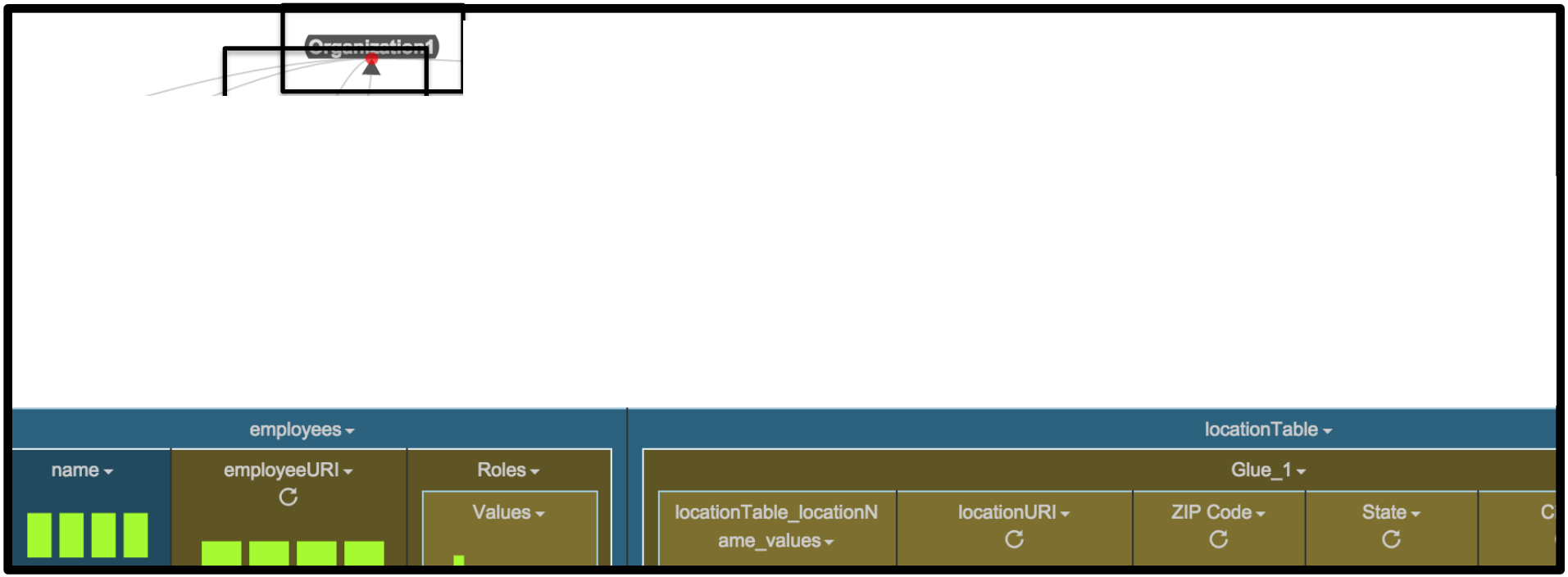


KR2RML applied to Nested Relational Model





KR2RML Processing



RDF Generation Triples Map Processing Order

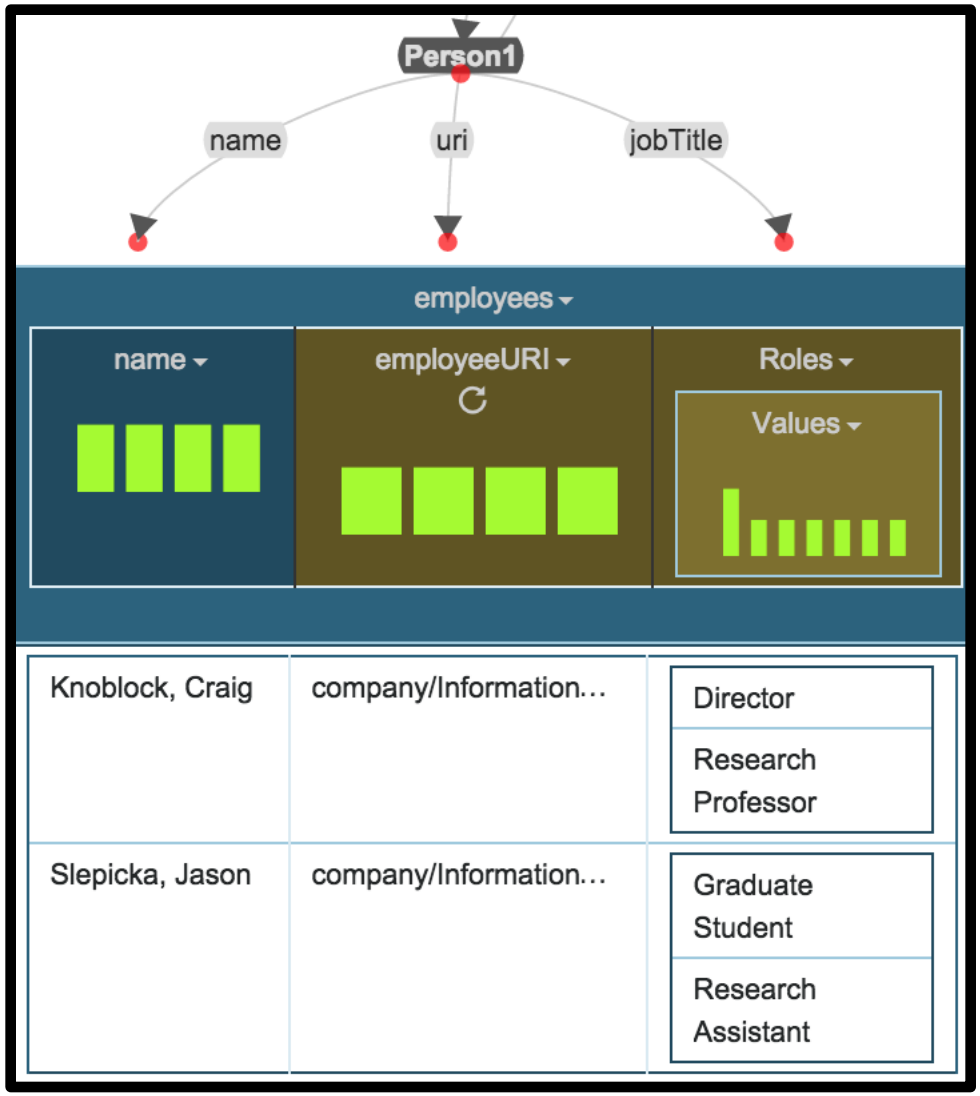
_:TriplesMap_4 (PostalAddress1)

_:TriplesMap_3 (Place1)

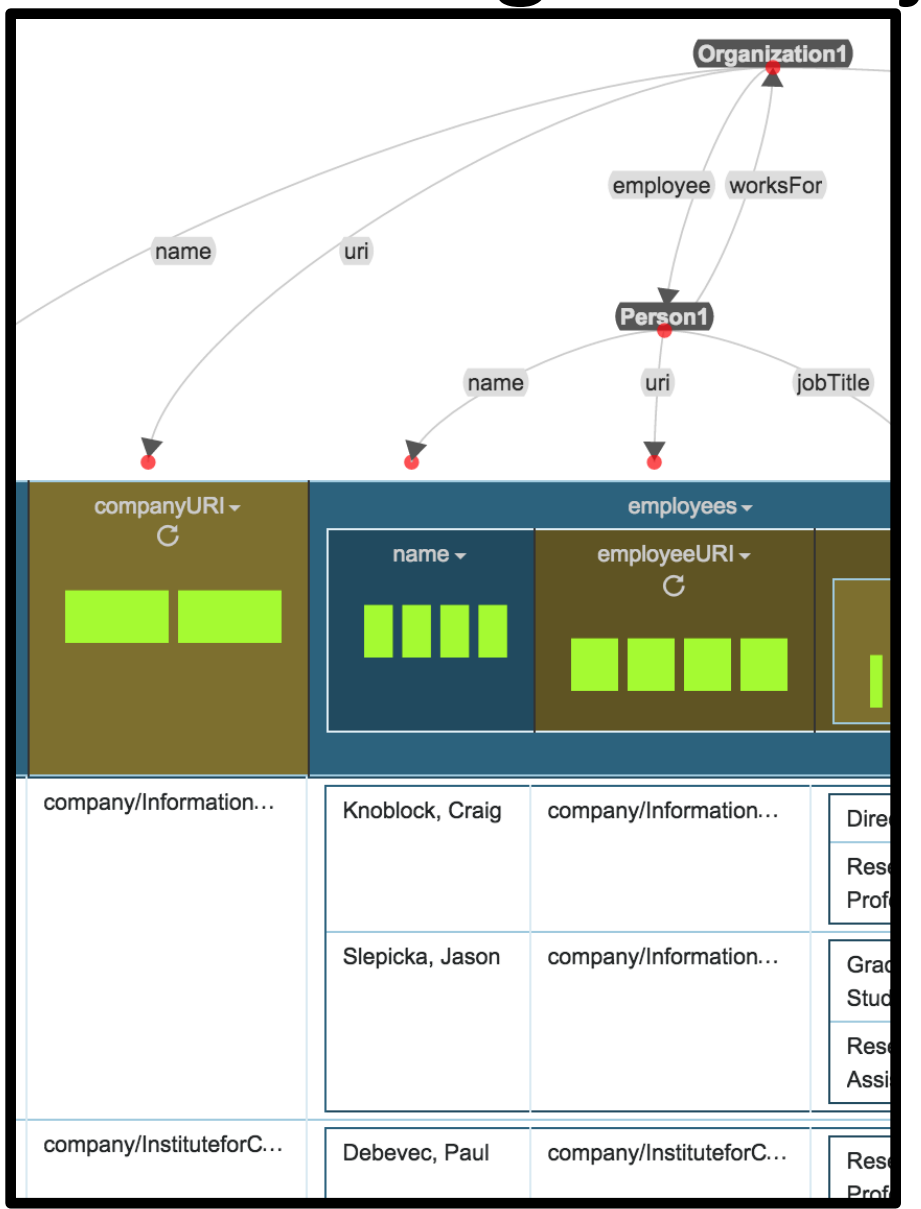
_:TriplesMap_2 (Person1)*

_:TriplesMap_1 (Organization1)

KR2RML Processing: ObjectMap



KR2RML Processing: RefObjectMap



KR2RML JSON-LD Output


```
{
 "@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?