OWL vs. Linked Data: Experiences and Directions

Axel Polleres
Siemens AG Österreich
27/05/2013
Will try to touch upon the following questions:

- Which parts of OWL are used within Linked Data?
- Which parts of OWL2 could be useful for Linked Data?
- Which reasoning techniques can be applied to handle the scale, messiness, and dynamicity of Linked Data?
- Which reasoning beyond RDFS and OWL is necessary for Linked Data?

I don’t have complete solutions to all these questions, but some on how OWL can be applied to Linked Data collected over the past view years.

Special thanks to my co-authors: Aidan Hogan, Jürgen Umbrich, Stefan Bischof, Andreas Harth, Birte Glimm, Markus Krötzsch, etc.
Let’s start from the beginning...

“If HTML and the Web made all the online documents look like one huge book, RDF, schema and inference languages will make all the data in the world look like one huge database”

Tim Berners-Lee, Weaving the Web, 1999

Great! We have 2013, this should work by now, shouldn’t it? Let’s try that on google!

**Scenario:** “Market research” on the Web, (no FOAF today, sorry)

```
Google

“Latest news on NYT about technology companies with a revenue greater than 10B EUR”

Google

“Which city of Montpellier and Vienna has the higher population density?”
```
“Latest News on NYT about technology companies with a revenue greater than 10B EUR”

• The data is there!

New York Times thesaurus and article search API
“Which city of Montpellier and Vienna has the higher population density?”

• Again:
  The data is there!

Urban Audit
(ca. 300 indicators for ~500 European cities)

or also inside Siemens!
Emerging trend: Linked data

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, OWL)
4. Include links to other URIs, so that they can discover more things.

http://www.w3.org/DesignIssues/LinkedData.html
Linked Data on the Web: Adoption
Example: Using OWL and SPARQL to query Linked data

Which city of Montpellier and Vienna has the higher population density?

```
SELECT ?DM ?DV
WHERE {
}
```
The Semantic promise...

“If HTML and the Web made all the online documents look like one huge book, RDF, schema and inference languages will make all the data in the world look like one huge database”

Tim Berners-Lee, Weaving the Web, 1999

Great! We have 2013, this should work by now, shouldn’t it? Let’s try that on google!

Scenario: “Market research” on the Web,

“Latest news on NYT about technology companies with a revenue greater than 10B EUR”

![Google search results]

```
SELECT * WHERE
  FILTER( ?R > 10000000000 ) } 
```
The Semantic promise...

“If HTML and the Web made all the online documents look like one huge book, RDF, schema and inference languages will make all the data in the world look like one huge database”

Tim Berners-Lee, Weaving the Web, 1999

Great! We have 2013, this should work by now, shouldn’t it? Let’s try that on google!

Scenario: “Market research” on the Web,

Google

“Latest news on NYT about technology companies with a revenue greater than 10B EUR”

Google

“Cities in France with a higher population density than Montpellier?”

```
SELECT ?DM ?DV
WHERE {
}
```
So, all we need to do:

- ... is to through together a crawler, an RDF Store, some OWL inference and a SPARQL engine? Unfortunately it’s not that easy…

```
SPARQL
SELECT ?DM ?DV WHERE
dbpedia:Vienna :populationDensity ?DV . }
```

So, all we need to do:
Obstacles/prejudices:

• OWL is too hard to learn?

• OWL is too expensive?
Linked Data publishers...

...OWL IS HARD

(...to learn, to understand, to implement, to compute, to teach, to represent in RDF, to publish, to parse, to use appropriately...)

| Corollary1: | Linked Data publishers only use a little bit of OWL … |
| Corollary2: | … they still manage to make mistakes 😊 |
Corollary 1: Linked Data publishers only use a little bit of OWL ...
• Looked at Billion Triple Challenge 2011 Dataset (BTC2011)
  – 2.1 billion quadruples, crawled from...
  – 7.4 million RDF/XML documents, covering...
  – 791 (pay-level) domains

• Count OWL features used in the dataset:
  – Per use
  – Per document
  – Per domain
  – Can be skewed by data

• Ranked OWL features using PageRank:
  – Rank documents based on dereferenceable links
  – For each OWL feature, sum the rank of documents using it
  – Intuition: Approximates probability of encountering an OWL feature during a random walk of the data
Results of ranking (see LDOW’12 paper details)

1. rdf:Property 5.74E-1
2. rdfs:range 4.67E-1
3. rdfs:domain 4.62E-1
4. rdfs:subClassOf 4.60E-1
5. rdfs:Class 4.45E-1
6. rdfs:subPropertyOf 2.35E-1
7. owl:Class 1.74E-1
8. owl:ObjectProperty 1.68E-1
9. rdfs:Datatype 1.68E-1
10. owl:DatatypeProperty 1.65E-1
11. owl:AnnotationProperty 1.60E-1
12. owl:FunctionalProperty 9.18E-2
13. owl:equivalentProperty 8.54E-2
14. owl:inverseOf 7.91E-2
15. owl:disjointWith 7.65E-2
### Results of ranking (see LDOW’12 paper details)

...  

<table>
<thead>
<tr>
<th>Property</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>owl:sameAs</td>
<td>7.29E-2</td>
</tr>
<tr>
<td>owl:equivalentClass</td>
<td>5.24E-2</td>
</tr>
<tr>
<td>owl:InverseFunctionalProperty</td>
<td>4.79E-2</td>
</tr>
<tr>
<td>owl:unionOf</td>
<td>3.15E-2</td>
</tr>
<tr>
<td>owl:SymmetricProperty</td>
<td>3.13E-2</td>
</tr>
<tr>
<td>owl:TransitiveProperty</td>
<td>2.98E-2</td>
</tr>
<tr>
<td>owl:someValuesFrom</td>
<td>2.13E-2</td>
</tr>
<tr>
<td>rdf:_*</td>
<td>1.42E-2</td>
</tr>
<tr>
<td>owl:allValuesFrom</td>
<td>2.98E-3</td>
</tr>
<tr>
<td>owl:minCardinality</td>
<td>2.43E-3</td>
</tr>
<tr>
<td>owl:maxCardinality</td>
<td>2.14E-3</td>
</tr>
<tr>
<td>owl:cardinality</td>
<td>1.75E-3</td>
</tr>
<tr>
<td>owl:oneOf</td>
<td>4.13E-4</td>
</tr>
<tr>
<td>owl:hasValue</td>
<td>3.91E-4</td>
</tr>
<tr>
<td>owl:intersectionOf</td>
<td>3.37E-4</td>
</tr>
<tr>
<td>owl:NamedIndividual</td>
<td>3.37E-4</td>
</tr>
</tbody>
</table>

Axel Polleres
How much OWL is used in Linked Data? [LDOW’12]

- RDF Schema features amongst the most prominently used
- OWL 2 features not yet used prominently
Main insight:

- RDF Schema features amongst the most prominently used
- OWL 2 features not yet used prominently
- Most *used* features can be implemented efficiently using parallelizable, rule-based inference. [IJSWIS’09]
Intuition: MAYBE LINKED DATA ONLY NEEDS A LITTLE OWL...

(...for now) [Hendler, ‘98]
How much OWL is used in Linked Data? [LDOW’12]

- RDF Schema features amongst the most prominently used
- OWL 2 features not yet used prominently

In all fairness:

This is only a snapshot...

... some OWL2 features could be quite useful to enrich current Linked Data vocabularies, see also [RR’09]
We presented this 3 years ago at ESWC... still only little OWL2 adoption in LOD vocabs since then...

Common ontologies on the Web don’t use it a lot as of yet...

... but adds interesting functionality, potentially useful for Web ontologies, e.g.

- PropertyChains
  - E.g. could be useful to tie sioc:name and foaf:nick via foaf:holdsAccount:

```
foaf:nick owl:propertyChainAxiom (foaf:holdsAccount sioc:name)
```

HOT! NEW!
Corollary 2: ... publishers still manage to make mistakes.

Examples...
4 main conjectures for possible reasons:

- Publishers deliberately publish spoilt data ("SPAM")
- Opinions differ
- "URI-sense" ambiguities
- **Accidentally** wrong/inconsistent
Publishers deliberately publish spoilt data (“SPAM”)

• Examples:
  
  – a owl:differentFrom a .
  
  – foaf:knows rdfs:subPropertyOf owl:sameAs .
  
  – http://www.polleres.net/nasty.rdf

• Can occur for “testdata” being published, deliberate SPAM might become an issue, as the SW grows!
Opinions differ

- **Fictitious Example Ontology:**
  
  `originofthings.example.org`:
  
  ```
  o1:surpremePower owl:disjointWith o1:naturalPhenomenom.
  o1:originsFrom rdf:type owl:functionalProperty.
  o1:god rdf:type o1:surpremePower.
  o1:evolution rdf:type o1:naturalPhenomenom.
  ```

  `darwin.example.org`:
  
  ```
  ex:mankind o1:originsFrom o1:evolution .
  ```

  `creationism.example.org`:
  
  ```
  ex:mankind o1:originsFrom o1:god
  ```

  `FlyingSpaghettimonster.org`
  
  ```
  ex:mankind o1:originsFrom fsm:theSpaghettiMonster.
  ```
"URI-sense" ambiguities

<http://www.polleres.net>

foaf:knows <http://apassant.net>

i.e., why do I have to use a different URI for myself and my homepage?

Many people don’t understand/like this and make mistakes.

But is this really a mistake or a design error?
Accidentally inconsistent data

This appeared a while ago in our crawls:

**Source1 (faulty):**
TimBL foaf:homepage <http://www.w3.org>
TimBL rdf:type foaf:Person.

**W3.org:**
W3C foaf:homepage <http://www.w3.org>
W3C rdf:type foaf:Organisation.

*Did occur in our Web crawls at some point, sometimes people don’t have the right semantics in mind!*

- Suspiciously resembles problems with e.g. flawed HTML ... browsers, normal search engines still have to deal with it...
Accidently wrong (non-inconsistent data)

• FOAF Ontology:

foaf:mbox rdf:type owl:InverseFunctionalProperty .

• Careless FOAF exporters produce something like this for any empty email address:
  ex:alice foaf:mbox “mailto:”
  ex:bob foaf:mbox “mailto:”
  …

Consequence: InverseFunctionalProperty reasoning on Web Data potentially equates too many things! Dangerous!

Not all of these were real examples, of course, for a more systematic analysis of Linked Data conformance, see [JWS’12a,LDOW’10]

BTW, Didn’t even mention ill-typed datatype literals here (very common...)
Your mission, should you decide to accept it, would be to make the Semantic Web clean ...

http://swse.deri.org/RDFAlerts/

URL: http://polleres.net/foaf.rdf

Please note that we currently only validate RDF/XML, and that validation may take 1-2 minutes due to live crawling of data.

| note  | error retrieving http://skype.com/ - http://skype.com/ did not return Content-Type application/rdf+xml |
| okay  | retrieved data                                                      |
| warning | could not find a definition for Class http://skype.com/... term does not dereference to an RDF vocabulary description? |
| warning | could not find a definition for Property http://xmlns.com/foaf/0.1/accountName... term does not dereference to an RDF vocabulary description |
| warning | could not find a definition for Property http://xmlns.com/foaf/0.1/accountName... term does not dereference to an RDF vocabulary description |
| warning | could not find a definition for Property http://xmlns.com/foaf/0.1/accountName... term does not dereference to an RDF vocabulary description |
| okay  | finished validation                                                |
Conclusion: This won’t scale to the Web...

```
SELECT ?DM ?DV WHERE {
}
```
How to possibly avoid nasty inferences?

SELECT ?DM ?DV
}
Scalabe Reasoning: Focus on Rules without ABox joins

**DL Syntax**

<table>
<thead>
<tr>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\forall C \sqsubseteq D \quad ?C \text{ rdfs:subClassOf } ?D, \quad ?s \text{ a } ?C \quad \Rightarrow \quad ?s \text{ a } ?D.$</td>
</tr>
<tr>
<td>$C \equiv D \quad \frac{?C \text{ :equivalentClass } ?D, \quad ?s \text{ a } ?C \quad \Rightarrow \quad ?s \text{ a } ?D.}{?C : \text{equivClass } ?D, \quad ?s \text{ a } ?C \quad \Rightarrow \quad ?s \text{ a } ?D.}$</td>
</tr>
<tr>
<td>$P \sqsubseteq Q \quad \frac{?P \text{ rdfs:subPropertyOf } ?Q, \quad ?s \text{ ?P }\text{ o} \quad \Rightarrow \quad ?s \text{ ?Q }?o.}{?P : \text{equivProperty } ?Q, \quad ?s \text{ ?P }\text{ o} \quad \Rightarrow \quad ?s \text{ ?Q }?o.}$</td>
</tr>
<tr>
<td>$P \equiv Q \quad \frac{?P : \text{equivProperty } ?Q, \quad ?s \text{ ?Q }?o \quad \Rightarrow \quad ?s \text{ ?P }?o.}{?P : \text{equivProperty } ?Q, \quad ?s \text{ ?Q }?o \quad \Rightarrow \quad ?s \text{ ?P }?o.}$</td>
</tr>
<tr>
<td>$\top \sqsubseteq \forall P.C \quad \frac{?P \text{ rdfs:domain } ?C, \quad ?s \text{ ?P }\text{ o} \quad \Rightarrow \quad ?s \text{ a } ?C.}{?P : \text{equivDomain } ?C, \quad ?s \text{ ?P }\text{ o} \quad \Rightarrow \quad ?s \text{ a } ?C.}$</td>
</tr>
<tr>
<td>$\top \sqsubseteq \forall P.C \quad \frac{?P \text{ rdfs:range } ?C, \quad ?s \text{ ?P }\text{ o} \quad \Rightarrow \quad ?s \text{ a } ?C.}{?P : \text{equivRange } ?C, \quad ?s \text{ ?P }\text{ o} \quad \Rightarrow \quad ?s \text{ a } ?C.}$</td>
</tr>
<tr>
<td>$P \equiv P' \quad \frac{?P : \text{hasValue } ?x; \quad \text{onProperty } ?P, \quad ?y \text{ ?P }?x \quad \Rightarrow \quad ?y \text{ a } ?C.}{?P : \text{hasValue } ?x; \quad \text{onProperty } ?P, \quad ?y \text{ ?P }?x \quad \Rightarrow \quad ?y \text{ ?P }?x.}$</td>
</tr>
<tr>
<td>$\exists P.x \quad \frac{?C : \text{unionOf } (?C_1, ..., ?C_n), \quad ?x \text{ a } ?C_i \quad \Rightarrow \quad ?x \text{ a } ?C.}{(\geq 1 P) \quad ?C : \text{minCardinality 1; } \text{onProperty } ?P, \quad ?x \text{ ?P }?y \quad \Rightarrow \quad ?x \text{ a } ?C.}$</td>
</tr>
<tr>
<td>$C_1 \sqsubseteq \ldots \sqsubseteq C_n \quad ?C : \text{intersectionOf } (?C_1, ..., ?C_n), \quad ?y \text{ a } ?C_i \quad \Rightarrow \quad ?y \text{ a } ?C_i, \quad ?y \text{ a } ?C, \quad \Rightarrow \quad ?y \text{ a } ?C_1, \quad \ldots, \quad ?y \text{ a } ?C_n, \quad \Rightarrow \quad ?y \text{ a } ?C.$</td>
</tr>
<tr>
<td>$C_1 \sqsubseteq \ldots \sqsubseteq C_n \quad ?C : \text{intersectionOf } (?C_1), \quad ?y \text{ a } ?C_1 \quad \Rightarrow \quad ?y \text{ a } ?C.$</td>
</tr>
</tbody>
</table>

From [IJSWIS’09] The rules applied including statements considered to be T-Box, elements which must be authoritatively spoken for (including for bnode OWL abstract syntax), and output count.

Abox inferences only! Can be distributed! Several Optimizations in [ISWC2010]
Authoritative Reasoning

- Document D authoritative for concept C iff:
  - C not identified by a URI
  - OR
  - De-referenced URI of C coincides with or redirects to D
- FOAF spec authoritative for foaf:Person ✓
- MY Ont not authoritative for foaf:Person ✗

- Only allow extension in authoritative documents
  - my:Person rdfs:subClassOf foaf:Person . (MY Ont) ✓
- BUT: Reduce obscure memberships
  - foaf:Person rdfs:subClassOf my:Person . (MY Ont) ✗
- Similarly for other T-Box statements.
- In-memory T-Box stores authoritative values for rule execution
A Semantic Web Search Engine

- ... based on these insights we have implemented SWSE [JWS’11]
Are we done?

Similar in spirit to SAOR: “Quarantined Reasoning”, cf. [RR’11]

“Which city of Montpellier and Vienna has the higher population density?”

“Latest news on NYT about technology companies with a revenue greater than 10B EUR”

SELECT * WHERE
FILTER( ?R > 10000000000 ) }
Idea: “Live”-Linked Data Querying

- Basic idea (originally proposed by Hartig & Bizer)
  - Start with URIs in a SPARQL query
  - Interleave crawl + query processing

```
HTTP GET NYT:Org  →  HTTP GET NYT:SAP  →  HTTP GET NYT:article20130128_1

SELECT * WHERE
  FILTER (?R > €10000000000 ) } 

Stop. No new query relevant results found
```
**Improved “Live”-Linked Date Querying**

- Our approach [RR’12]
- Basic idea:
  - Start with URIs in a SPARQL query
  - Interleave crawl + query processing
  + Lightweight OWL Reasoning

```
SELECT * WHERE 
  ?C NYT:latestArticle ?A.
  FILTER (?R > €10000000000) }
```

HTTP GET NYT:Org → HTTP GET NYT:SAP → HTTP GET dbpedia:SAP
Again: Does this work “in the wild”?

1) Benchmark generation:
   - BTC 2011
   - QWalk: Random walk based query generation.
   - 1100 queries
     - 100 each for 11 “typical” shapes

2) Results:
   - **RDFS** reasoning: many queries with average impact
     - In 8/11 query classes more than 50% avg. result increase
     - In 4/11 query classes a time increase of 50%
   - **OWL** sameAs: view queries with high impact
     - In 2/11 query classes more than 50% avg. result increase
     - In 2/11 query classes a time increase of 50%

Overall our reasoning extensions improved in 8 of 11 query classes the average result/time ratio (throughput)
Are we done?

Which city of Montpellier and Vienna has the higher population density?

Latest news on NYT about technology companies with a revenue greater than 10B EUR

Depends on the point of view…
... e.g.: What is the sweetspot between **centralised & decentralised/live** Linked Data querying approaches?

Plus…
... Actually, we were cheating:

```
SELECT ?DM ?DV
}
```

“Which city of Montpellier and Vienna has the higher population density?”
Example: Actually, we were cheating…

Data looks rather like this…

“Which city of Montpellier and Vienna has the higher population density?”

Can’t we still answer this query? Just using the Dbpedia data?

We know: \( \text{dbpedia:populationDensity} = \frac{\text{dbpedia:population}}{\text{dbpedia:area}} \)

Another example is unit conversion from this morning:

\( \text{elephantWeight} = \text{elephantWeightImperial} \times 2.204 \)

or

\( \text{priceUSD} = \text{priceEUR} \times 1.2850 \)

We call these “property equation” … stay tuned for the main track [ESWC’13]
Take-home messages:
Lots of interesting real-world Datasets out there as Linked Data!

- can benefit from OWL and Reasoning
- (even more Open Data in other formats, BTW)

Structured Query answering over Linked Data needs:

- **Robust & Scalable Reasoning** (à la SAOR)
- Both **centralised & decentralised** approaches
- Maybe more than OWL (e.g. something like property equations)?

Vienna

<table>
<thead>
<tr>
<th>Year</th>
<th>Density</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>4002.2 pl/km²</td>
<td><a href="http://dctpedia.org">http://dctpedia.org</a></td>
</tr>
<tr>
<td>2004</td>
<td>3885.25 pl/km²</td>
<td><a href="http://www.urbanaudit.org">http://www.urbanaudit.org</a></td>
</tr>
<tr>
<td>2001</td>
<td>3735.24 pl/km²</td>
<td><a href="http://www.urbanaudit.org">http://www.urbanaudit.org</a></td>
</tr>
<tr>
<td>1999</td>
<td>3735.2 pl/km²</td>
<td><a href="http://eurostat.linked-statistics.org">http://eurostat.linked-statistics.org</a></td>
</tr>
</tbody>
</table>
References:


[RR’09] Aidan Hogan, Stefan Decker: On the Ostensibly Silent 'W' in OWL 2 RL. *RR 2009*: 118-134


...(more on www.polleres.net/publications.html)
Open Data Trends & Future

- Open Data: Typically very liberal licenses (variants of CC)
- Many formats, varying quality, harmonization starting
- Mostly by online communities or public bodies (cities, communities, governments, UN, ...)
  - Currently focused mostly in SMEs to take advantage of that data
- vs. Publicly available data: e.g. NYT is public but not free/not license free
- vs. Enterprise (Linked) Data

Axel Polleres
Linked Data Ontologies = RDF Vocabularies (OWL, RDFS)

Side remark: Only a small fraction of Linked data is OWL/RDFS ontologies (less than ~0.1% [IJSWIS’2009, Hogan 2010])
OWL Reasoning: Complexity


And even for OWL RL PTIME-complete.... i.e., often too expensive.
Related Works:

- **Scalable OWL Reasoning**

- **Live Linked Data Querying**