Weaving the Pedantic Web

LDOW 2010
Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, Axel Polleres
Linked Data...
Purpose of talk: Application developers... how to not sink...
Purpose of talk: *RDF Publishers... how to avoid common mistakes...*
Talking about errors in Linked Data...

We’ll try not to ruin the party

...statistics based on crawl:
- April 2009
- 5k domain limit
- 150k URIS, 55k RDF docs
- 12.5m triples (quads)
- Mentioning 1.6m URIs
- 5,850 classes/9,507 props
- Accept: application/rdf+xml

...okay... so no RDFa

Statistics are *illustrative*
Chapter 1: HTTP-level issues...

...a good RDF description these days is hard to find
Waldo URIs.

**URLs with no dereferencable RDF**

**Not a crawler’s idea of fun...**
Hmm not *so* many...

- 5.3% of HTTP URIs return 40x/50x
- Excluding redirects...
  92.8% return 200 OK

- In return, only 45.4% of 200 Okay return report application/rdf+xml
- 34.8% return HTML... probably just HTML docs...
  okay... maybe a *few* contain RDFa
Lies... Damned Lies... 
& **Content-Type Reporting**

“Trust me, it’s RDF/XML”
Okay... So he’s actually pretty honest

- **16.9% of valid RDF/XML documents returned with an invalid/more generic Content-type:**
  - text/xml (9.5%)
  - application/xml (5.9%)
  - text/plain (1%)
  - text/html (0.4%)

- **Of those returning Content-type:application/rdf+xml**
  - 98.8% were valid RDF/XML
I wish they’d used a redirect...
E.g., the **Miracle at Calais:**

*turning 1,778 triples into $\sim \infty$ quads*

http://d.opencalais.com/1/type/em/r/SameTriplesDifferentDocument

*(apologies to OpenCalais guys - it’s just a convenient example)*
Chapter 2: Reasoning issues...

...or, how I learned to start worrying and stop loving OWL
Undefined classes and properties...

It looks important, but I’m afraid I don’t fully follow...
Quite common...

- 14.3% of triples use undeclared property
- 8.1% of triples use undeclared class

Three cases:

- Case 1: *Namespace has no vocabulary/is not deferencable*  
  (e.g., `rss:item`)

- Case 2: *Term invented in related namespace*  
  (e.g., `foaf:tagLine` invented by LiveJournal)

- Case 3: *Term is misspelt version of term defined in namespace*  
  (e.g., `foaf:image` vs. `foaf:img`)
Despite what you claim, not all of you can *actually be* Spartacus
Spartacus relived...

08445a31a78661b5c746feff39a9db6e4e2cc5cf

- **sha1-sum of `mailto:`**
- **common value for foaf:mbox_sha1sum**
  - An inverse-functional (uniquely identifying) property!!!
  - Any person who shares the same value will be considered the same

*I’m Spartacus!*...and so’s my wife
As he would undoubtedly be able to tell you, “true” is not a valid xsd:int
Not *too* bad…

- 4.7% of typed literals were “ill-typed” (lexically invalid)...
  - mostly `xsd:dateTime` (26.4% of all date-time literals were invalid; e.g., omitted the seconds field)

- Also, literals are sometimes incompatible with the datatype-range of a property:
  - E.g., 21.8% of `ical:description` triples used language tags incompatible with the defined range of `xsd:string`
  - E.g., 100% of `sl:creationDate` triples use plain literal values incompatible with defined range of `xsd:date`
Mystical beings...

**Members of disjoint classes**

Despite what FOAF says, it seems that **Persons can also be Documents**
Again, not *too* bad…

- 1,329 members of disjoint classes found

- Generally caused by naïve URI naming:
  - Use of information resource URIs to name entities (particularly foaf:Persons)
  - E.g., `<me> foaf:knows <jim/foaf.rdf>`.
Ontology hijacking...

Anybody can say anything, anywhere, and unfortunately for everyone else, have a good chance of being taken seriously.
Redefining Everything...  
...and home in time for tea

From http://www.eiao.net/rdf/1.0

<owl:Property rdf:about="http://www.w3.org/1999/02/22-rdf-syntax-ns#type">  
  <rdfs:label xml:lang="en">type</rdfs:label>  
  <rdfs:comment xml:lang="en">Type of resource</rdfs:comment>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#testRun"/>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#pageSurvey"/>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#siteSurvey"/>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#scenario"/>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#rangeLocation"/>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#startPointer"/>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#endPointer"/>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#header"/>  
  <rdfs:domain rdf:resource="http://www.eiao.net/rdf/1.0#runs"/>  
</owl:Property>

Ontology hijacking!!  
(apologies to EIAO guys - it’s just a convenient example)
Solutions?
Application side: workarounds

- All presented issues have a suitable antidote, *once you know about them*

- See paper for discussion...
Syntax errors quite rare, partly due to popularity of W3C RDF/XML syntax validator

Need an all-in-one validation service

- Should not only validate strict errors, but give feedback on suspected issues
- We offer a prototypical service at: http://swse.deri.org/RDFAlerts/
- Get the community to contact publishers about errors/issues as they arise
- Get involved: http://pedantic-web.org/
- 137 members!

Acknowledgements to: Aidan Hogan, Alex Passant, Me, Antoine Zimmermann, Axel Polleres, Michael Hausenblas, Richard Cyganiak, Stéphane Corlosquet
...unattended, can be pretty serious...

\[
\text{foaf:mbox_sha1sum} \ a \ \text{owl:InverseFunctionalProperty} . \\
?x \ \text{foaf:mbox_sha1sum} \ 08445a31a78661b5c746feff39a9db6e4e2cc5cf . \\
\]

**OWL 2 RL rule prp-ifp:**

\[
?p \ a \ \text{owl:InverseFunctionalProperty} . \ ?x_1 \ ?p \ ?z . \ ?x_2 \ ?p \ ?z . \\
\Rightarrow \ ?x_1 \ \text{owl:sameAs} \ ?x_2 .
\]

10^6 \ ?x_1/?x_2 \ binding\s in body

\Rightarrow \ 10^{12} \ inferred \ pair-wise \ and \ reflexive \ \text{owl:sameAs} \ statements

...or in simpler terms: