RDFa: Embedding RDF Knowledge in HTML

Some content from a presentation by Ivan Herman of the W3c, [Introduction to RDFa](#), given at the 2011 Semantic Technologies Conference.
What is RDFa?

- Serialization of RDF embedded in HTML, HTML or XML
  Provides set of attributes (the `a` in RDFa) to use with existing tags to carry RDF metadata
- 2004: work on developing standards began
- 2008: RDFa 1.0 a recommendation (but only in XHTML, which failed to launch)
- 2012-15: RDFa 1.1 recommendation (works in HTML4, HTML5 and XHTML)
- See [http://rdfa.info/](http://rdfa.info/)
Principles of RDFa

- RDF content specified in XML *attributes* of tags rather than *elements*
- The XML/HTML *tree structure* is used as context, when appropriate
- Some new attributes are *introduced* and some existing ones (@href, @rel) *reused*
- When possible, HTML text content used for *literal values*

→ Same file used by browser & RDF extractor
"http://www.w3.org/ns/entailment/RDFS" is the URI. The specification for the RDFS entailment is part of the RDF Semantics W3C Recommendation.

For more information about RDF, please refer to the the RDF Concepts and Abstract Syntax Recommendation.

Ivan Herman, ivan@w3.org, W3C, Semantic Web Activity Lead, 2009-05-03
The source

<p about="http://www.w3.org/ns/entailment/RDFS"
    property="http://purl.org/dc/terms/description">
    Unique identifier for <em>RDFS Entailment</em>.
</p>
Unique identifier for <em>RDFS Entailment</em>.
Unique identifier for <em>RDFS Entailment</em>.
Unique identifier for RDFS Entailment.
Unique identifier for RDFS Entailment.

“http://www.w3.org/ns/entailment/RDFS” is the URI. The specification for the RDFS entailment is part of the RDF Semantics W3C Recommendation.

For more information about RDF, please refer to the RDF Concepts and Abstract Syntax Recommendation.

Ivan Herman, ivan@w3.org, W3C, Semantic Web Activity Lead, 2009-05-03
<a about="http://www.w3.org/ns/entailment/RDFS"
    rel="http://www.w3.org/2000/01/rdf-schema#seeAlso"
    href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
    RDF Semantics.
</a>
Source and generated RDF...

```html
<a about="http://www.w3.org/ns/entailment/RDFS"
   rel="http://www.w3.org/2000/01/rdf-schema#seeAlso"
   href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/>
   RDF Semantics.
</a>

<http://www.w3.org/ns/entailment/RDFS>
   ....
```
Source and generated RDF...

<a about="http://www.w3.org/ns/entailment/RDFS" rel="http://www.w3.org/2000/01/rdf-schema#seeAlso" href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
   RDF Semantics.
</a>
Source and generated RDF...

```xml
<a about="http://www.w3.org/ns/entailment/RDFS"
   rel="http://www.w3.org/2000/01/rdf-schema#seeAlso"
   href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
   RDF Semantics.
</a>
```

```xml
<http://www.w3.org/ns/entailment/RDFS>
   <http://www.w3.org/2000/01/rdf-schema#seeAlso>
```
Ntriples in HTML 😞

- Maybe we can do better, instead of this

```
<http://www.w3.org/ns/entailment/RDFS>
  <http://purl.org/dc/terms/description>
    "Unique identifier for RDFS Entailment." .
<http://www.w3.org/ns/entailment/RDFS>
  <http://www.w3.org/2000/01/rdf-schema#seeAlso>
```

- Allow URI prefixes and shared subject, like this

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .

<http://www.w3.org/ns/entailment/RDFS>
  rdfs:seeAlso <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> ;
  dcterms:description "Unique identifier for RDFS Entailment." .
```
Turtlizing RDFa

- Turtle supports several simplifying ideas
- Use **compact URIs** *(CURIE)* when possible
  - URI with a prefix defined elsewhere, e.g., *foaf:mbox*
- Making use of the natural structure for
  - shared subjects
  - shared predicates
  - creating blank nodes
  - etc.
CURIE definition and usage

• can be replaced by:

```html
<html prefix="dcterm: http://purl.org/dc/terms/">
  ...
  <p about="http://www.w3.org/ns/entailment/RDFS"
      property="dcterm:description">
    Unique identifier for <em>RDFS Entailment</em>.</p>
  ...
</html>
```
Details on @prefix in RDFa

- Can be anywhere in the HTML tree and is valid for whole sub-tree
  - i.e., html element is not the only place to have it
- The same @prefix attribute can hold several definitions:
  - prefix="dcterm: http://purl.org... foaf: http://..."
- CURIEs and “real” URIs can usually be mixed
- CURIEs cannot be used on @href
Basic principle: @about is inherited by children nodes, so no reason to repeat it

```html
<html prefix="dcterms: http://purl.org/dc/terms/
                   rdfs: http://www.w3.org/2000/01/rdf-schema#">

...<body about="http://www.w3.org/ns/entailment/RDFS">

...<p property="dcterms:description">
  Unique identifier for <em>RDFS Entailment</em>.</p>
<p>...<a rel="rdfs:seeAlso" href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210">
  RDFS Semantics</a>...</p>
```
... yielding

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .

<http://www.w3.org/ns/entailment/RDFS>
    rdfs:seeAlso <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> ;
    dcterms:description "Unique identifier for RDFS Entailment." .
On reusing literals

- Reusing literals is a plus, but you don’t always want to do it
- The basic rule says: the (RDF) Literal is the enclosed text from the HTML content
- This is fine in 80% of the cases, but...
- ...it may not be natural in many cases!
Example: dates

```xml
<body about=".." prefix="dcterms: http://... xsd: http://..."

  <address>
    <p property="dcterms:date" datatype="xsd:date">2010-07-05</p>
  </address>
</body>
```

- This leads to:

```prolog
@prefix dcterms: <http://...> .
@prefix xsd: <http://...> .
<..> dcterms:date "2010-07-05"^^xsd:date .
```

- 2010-07-05 is the official ISO format (for xsd:date) but “July 5, 2010” is preferred by people
Usage of @content

Also leads to:

@prefix dcterms: <http://...> .
@prefix xsd: <http://...> .
<..> dcterms:date "2010-07-05"^^xsd:date .
On subjects and objects

- Here is our rule so far
  - @about sets the subject
  - @href sets the object

- But that is not always good enough
  - We may not want to introduce an active link (i.e., "a" element) on the web page
  - what about other links in HTML?
We may not always want links...

- The RDFa @resource attribute is equivalent to @href
- Sets the object, just like @href but is ignored by browsers, e.g.,:

```html
<span about="http://www.ivan-herman.net/foaf#me">
  <span rel="rdfs:seeAlso"
    resource="http://www.w3.org/People/Ivan/">Activity Lead</span>
</span>
```
More features

- RDFa1.1 has more features that make it easier to represent knowledge compactly in HTML.
- These take advantage of the HTML tree context.
- We’ll skip these details, which you can find in
  - RDFa 1.1 Primer
  - RDFa 1.1 Core
Some tools already have RDFa facilities:
- e.g., it is possible to add the right DTD to Dreamweaver, Amaya has it at its core, etc.

There are plugins to, e.g., WordPress, to generate RDFa markup

CMS systems (like Drupal 7) may have RDFa built in their publication system
- users generate RDFa whether they know about it or not...
Consuming RDFa

- Major search engines (Google, Yahoo) process RDFa for vocabularies they understand can use
- There are libraries, distillers, etc., to extract RDFa information
  - may be part of RDF development environments like Redland, RDFLib
  - see, for further references, http://rdfa.info/wiki/Consume
- Facebook’s “social graph” is based on RDFa
Publishing RDFa

- RDFa+HTML file can just be on a server
  - the client extracts the RDF content
- Content negotiations can be set up on the server side
  - the client gets the format he/she asks for
  - the RDF content can either be generated on the fly or stored on the server statically
Google’s rich snippets

Embedded metadata (microdata or RDFa) is used to improve search result page
- at the moment only a few vocabularies are recognized, but that will evolve over the years
A number of popular sites publish RDFa as part of their normal pages:

- Tesco, BestBuy, Slideshare, The London Gazette, Newsweek, MSNBC, O’Reilly Catalog, the White House...
- Creative Commons snippets are in RDFa (e.g., on Flickr)
BestBuy example of RDFa use
BestBuy example of RDFa Use

Courtesy of Jay Myers, BestBuy, SemTech2010 Presentation
Effects on BestBuy

- Reported in a BestBuy blog:
  - GoodRelations+RDFa improved Google rank tremendously
  - 30% increase in traffic on BestBuy store pages
  - Yahoo observers a 15% increase in click-through rate

- Today, BestBuy uses RDFa for much more than just snippets
  - E.g., to locate shops that have certain products on stock...
## Semantic Web

### From Library of Congress Subject Headings

<table>
<thead>
<tr>
<th>Details</th>
<th>Visualization</th>
<th>Suggest Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Semantic Web</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>URI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>&lt;http://id.loc.gov/authorities/sh2002000569#concap&gt;</code></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topical Term</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Broader Terms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic integration (Computer systems)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic networks (Information theory)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World Wide Web</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Related Terms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microformats</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sources</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engr. index online, May 6, 2002 (identifier: Semantic Web)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASTI on FirstSearch, May 6, 2002: in titles (semantic Web)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LC Classification</strong></td>
<td>TK5105.88815</td>
<td></td>
</tr>
</tbody>
</table>
<p rel="skos:inScheme" resource="http://id.loc.gov/authorities#topicalTerms">Topical Term</p>

<li>
</li>

<li>
<h3>Broader Terms</h3>
<ul class="std">
  <li>
    <a href="http://id.loc.gov/authorities/sh200400479#concept" rel="skos:broader">Semantic integration (Computer systems)</a>
  </li>
  <li>
    <a href="http://id.loc.gov/authorities/sh920049149#concept" rel="skos:broader">Semantic networks (Information theory)</a>
  </li>
  <li>
    <a href="http://id.loc.gov/authorities/sh950005419#concept" rel="skos:broader">World Wide Web</a>
  </li>
</ul>
</li>
Overstock.com example
gr:hasBusinessFunction gr:Sell;
gr:hasPriceSpecification [ gr:hasCurrency "USD"^^xsd:string;
gr:hasCurrencyValue "68.99"^^xsd:float ];
Drupal content management system

- RDF support in Drupal v. 7
- Major CMS system
- Has RDF at his core, pages contain RDFa
- In one step millions of pages of additional RDF data!
Sheriff Dupnik names Angle and Palin as two responsible for 'vitriol' (Video)

Yesterday Pima County Sheriff Charles Dupnik took the bold step of naming Sharon Angle and Sarah Palin as two individuals responsible for the kind of toxic 'political vitriol' which he believes must end in light of the recent shooting in Tucson, Arizona. Dupnik's comments came in an interview with Geraldo Rivera of Fox News. Earlier in the day Dupnik had been interview by Megyn Kelly of Fox News, who accused the Sheriff of inserting politics where it did not belong. In the Rivera interview, Dupnik stands by his earlier comments, and goes further in identifying two Republicans as the source...
Extracting the data

```
rdfa> python getdata.py "http://www.w3.org/ns/entailment/data/RDFS.html"
@prefix dc: <http://purl.org/dc/terms/> .
@prefix ent: <http://www.w3.org/ns/entailment/> .
...
ent:RDFS a ent:Entailment ;
    dc:creator <http://www.ivan-herman.net/foaf#me> ;
    dc:date "2010-05-03"^^xsd:date ;
    dc:description "Unique identifier for RDFS Entailment" ;
    rdfs:comment "The specification for the RDFS entailment is ... Semantics W3C Recommendation." ;
    rdfs:isDefinedBy <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/#rdfs_entailment> ;
    rdfs:seeAlso <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> .
<http://www.w3.org/ns/entailment/data/RDFS.html> dc:title "Information Resource RDFS Entailment" ;
    xhv:stylesheet <http://www.w3.org/StyleSheets/TR/base> .
<http://www.ivan-herman.net/foaf#me> a foaf:Person ;
    rdfs:seeAlso <http://www.ivan-herman.net/foaf> ;
    foaf:mbox <mailto:ivan@w3.org> ;
    foaf:name "Ivan Herman" ;
    foaf:title "Semantic Web Activity Lead" ;
    foaf:workplaceHomepage <http://www.w3.org> .
```
import rdflib, sys

if not (1 < len(sys.argv) < 4):
    print 'usage: python getdata.py url [ rdfa | rdfa1.1 | microdata | html ]'
    print ' eg: python getdata.py "http://www.w3.org/ns/entailment/data/RDFS.html"
    sys.exit(0)

url = sys.argv[1]
format = sys.argv[2] if len(sys.argv) == 3 else 'rdfa1.1'
g = rdflib.Graph()
g.parse(url, format=format)
print g.serialize(format='n3')
OpenLink Structured Data Sniffer

* http://osds.openlinksw.com/
OpenLink Structured Data Sniffer

https://osds.openlinksw.com/
Conclusions

- Web developers want content providers to add structured data to HTML pages.
- Content providers are incentivised to do so because their content will be better understood, ranked higher, more useful, etc.
- RDFa is most powerful & flexible knowledge markup standard understood by search engines.
- RDFa is also an alternative serialization of full RDF.