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.

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 is used for literal values
  ➔ Same file used by browser & RDF extractor

What is RDFa?

- Simple idea: a serialization of RDF embedded in XHTML, HTML or XML
  Provides a set of attributes (the a in RDFa) to use with existing tags to carry the RDF metadata
- 2004: work on developing standards began
- 2008: RDFa 1.0 a recommendation
  - Worked only in XHTML, which did not catch on
- 2012: RDFa 1.1 a recommendation
  - Works in HTML4, HTML5 and XHTML
- See http://rdfa.info/

Web page viewed by a person

http://www.w3.org/ns/entailment/data/RDFS.html
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>

Source and generated RDF...

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

Source and generated RDF...

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

Source and generated RDF...

<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 RDF Entailment."
The Web page viewed by a person

Unique identifier for RDF Entailment.

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

The source

`<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...

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

Turtlizing RDFa

- Turtle supports several simplifying ideas
- Use compact URIs when possible
  - A CURIE or compact URI, typically a URI with a prefix defined elsewhere, e.g., foaf:mbox
- Making use of the natural structure for
  - shared subjects
  - shared predicates
  - creating blank nodes
  - ...

We have Ntriples in HTML

- Maybe we can do better, instead of this

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

- Allow URI prefixes and shared subject, like this

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

</html>
```

CURIE definition and usage

- Can be replaced by:

```html
PREFIX dcterms: http://purl.org/dc/terms/

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

- Can be anywhere in the tree and is valid for the whole sub-tree
  - i.e., html element is not the only place to have it
- The same @prefix attribute can hold several definitions:
  - prefix="dcterms: http://purl.org... foaf: http://..."
- CURIEs and “real” URIs can usually be mixed
- CURIEs cannot be used on @href

Sharing subjects

The basic principle: @about is inherited by children nodes, so there’s 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 &lt;em&gt;RDFS Entailment&lt;/em&gt;.&lt;/p&gt;
    <p>&lt;a rel="rdfs:seeAlso" href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210">
      RDFS Semantics&lt;/a&gt;.&lt;/p&gt;
  </body>
</html>
```

... yielding

```html
@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

```
例：
    例
  > dates
  π prefix="dcterms: http://... xsd: http://..."
  <address>
    <p property="dcterms:date" datatype="xsd:date">2010-07-05</p>
  </address>
</body>
```

This leads to:

```
@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

```
例：
    例
  > dates
  π prefix="dcterms: http://... xsd: http://..."
  <address>
    <p property="dcterms:date" datatype="xsd:date" content="2010-07-05">July 5, 2010</p>
  </address>
</body>
```

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.,:

```
例：
    例
  > dates
  π about="http://www.ivan-herman.net/foaf#me" 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
- See the hidden slides if you are interested

Chaining

- A straightforward way:

```html
<body about="http://www.w3.org/ns/entailment/RDFS">
  ...
  <address>
    <span rel="dcterms:creator" resource="http://www.ivan-herman.net/foaf#me"/>
    <a rel="foaf:mailbox" href="mailto:ivan@w3.org">ivan@w3.org</a>,
    <a rel="foaf:workplaceHomepage" href="http://www.w3.org">W3C</a>
  </span>
</address>
```

Chaining

- A straightforward way:

```html
<body about="http://www.w3.org/ns/entailment/RDFS">
  ...
  <address>
    <span rel="dcterms:creator" resource="http://www.ivan-herman.net/foaf#me"/>
    <a rel="foaf:mailbox" href="mailto:ivan@w3.org">ivan@w3.org</a>,
    <a rel="foaf:workplaceHomepage" href="http://www.w3.org">W3C</a>
  </span>
</address>
```
Chaining: when objects become subjects

- An alternative:

```html
<body about="http://www.w3.org/ns/entailment/RDFS">
  ...
  <address>
    <span rel="dcterms:creator" resource="http://www.ivan-herman.net/foaf#me">
      <a rel="foaf:mailbox" href="mailto:ivan@w3.org">ivan@w3.org</a>,
      <a rel="foaf:workplaceHomepage" href="http://www.w3.org">W3C</a>
    </span>
  </address>
</body>
```

Chaining means

- @resource (or @href) becomes a subject for the sub-tree
- This feature is a bit like in RDF/XML

Some extra features

- Blank nodes can be created using "_:XX"
- Shorthand for RDF types
- Helping single-vocabulary cases
- Profiles

Typing

- Typing can of course be done using @rel="rdf:type"
- But that is a widely used combination, so there is a separate @typeof attribute for that
Typing example

```xml
<span about="http://www.ivan-herman.net/foaf#me" typeof="foaf:Person">
  <span property="foaf:name">Ivan Herman</span>,
</span>,

yields

```xml
http://www.ivan-herman.net/foaf#me a foaf:Person ;
  foaf:name "Ivan Herman" .
```

Single-vocabulary case

- In many cases the content is dominated by one vocabulary, e.g., dcterms, foaf, etc.
- CURIEs and URIs use is intuitive for RDF people but not for average HTML authors!
- Solution:
  - define a vocabulary URI for a sub-tree
  - for that sub-tree, simple terms in @rel, @property, etc., are automatically expanded into a full URI using the vocabulary

@vocab and terms: this...

```xml
<address about="http://www.ivan-herman.net/foaf#me" typeof="Person">
  <span property="foaf:name">Ivan Herman</span>,
  <a rel="foaf:mailbox" href="mailto:ivan@w3.org">ivan@w3.org</a>,
  <a rel="foaf:workplaceHomepage" href="http://www.w3.org">W3C</a>,
</address>
```

...becomes

```xml
<address about="http://www.ivan-herman.net/foaf#me" typeof="Person">
  <span property="name">Ivan Herman</span>,
  <a rel="mailbox" href="mailto:ivan@w3.org">ivan@w3.org</a>,
  <a rel="workplaceHomepage" href="http://www.w3.org">W3C</a>,
</address>
```
Prefix and term declarations can be collected in a separate file and referred to via a @profile attribute pointing to the file

Say, file “http://ex.org/prof” defines
  - prefix mappings:
    - "foaf" → "http://xmlns.com/foaf/0.1/"
    - "rdfs" → http://www.w3.org/2000/01/rdf-schema#
  - term mapping:
    - "desc" → http://purl.org/dc/terms/description
...becomes

```html
<html profile="http://ex.org/prof">
  ...
  <body about="http://www.w3.org/ns/entailment/RDFS">
    ...
    <p property="desc">
      Unique identifier for <em>RDFS Entailment</em>.</p>
    ...
    <address about="http://www.ivan-herman.net/foaf#me">
      <span property="foaf:name">Ivan Herman</span>,
      ...
    
    <p rel="rdfs:seeAlso" href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210">
      RDFS Semantics</a>.</p>
    ...
  </body>
  ...
</html>
```

Default profiles

- Even usage of profiles might be “too much” for many HTML authors 😞
  - authors will forget to add the @profile declaration
- RDFa defines default profiles:
  - RDFa clients include these profiles automatically

Default profiles

- Default for RDFa in general
  - [http://www.w3.org/profile/rdfa-1.1](http://www.w3.org/profile/rdfa-1.1)
  - includes some widely used prefixes (rdf, rdfs, vcard, og, foaf, dc, or dcterms are typical candidates)
  - the profile is to be updated regularly by adding new prefixes
- Default for (X)HTML
  - [http://www.w3.org/profile/html-rdфа-1.1](http://www.w3.org/profile/html-rdфа-1.1)
  - includes the HTML4 @rel values (next, up, license, ...)
  - the profile is to be updated regularly by adding @rel values as they evolve in the HTML world
So this...

```
<html profile="http://ex.org/prof">
  <body about="http://www.w3.org/ns/entailment/RDFS">
    <p property="dcterms:description">
      Unique identifier for <em>RDFS Entailment</em>.
    </p>
    <p rel="rdfs:seeAlso" href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210">
      RDFS Semantics</a></p>
  </body>
</html>
```

...becomes:

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

Authoring RDFa

- 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

Effects of, e.g., Google of Facebook

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 xxample of RDFa use
BestBuy example of RDFa Use

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...

Library of Congress RDFa use

Semantic Web
From Library of Congress Subject Headings

- Type: Semantic Web
- URI: http://id.loc.gov/authorities/subjects/sh85052157
- Related Terms:
  - Semantic Web
  - RDFa
- Source:
  - Date: 2002

Library of Congress RDFa use

Semantic Web
Overstock.com example

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!

The Examiner.com
ExtracLng the data

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')
Greenturtle Chrome plugin

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 the most powerful and flexible of the knowledge mark up standards understood by search engines
- RDFa is also an alternative serialization of full RDF