Embedding Knowledge in HTML

Some content from a presentations by Ivan Herman of the W3c

- We usually think of HTML as the language of Web pages
- But it’s also widely used on/for mobile devices and tablets
  - It readily adapts for different screen sizes/orientations
- And is the basis of many ebook formats
  - E.g. Kindle
- How can we add knowledge to HTML pages?

Adding RDF-like data to HTML

- We’d like to add semi-structured knowledge to a conventional HTML document
  - Humans can see and understand the regular HTML content (text, images, videos, audio)
  - Machines can see and understand the data markup in XML, RDF or some other format
- Possibilities include
  - Add a link to a separate document with the knowledge
  - Embed the knowledge as comments, javascript, etc.
  - Distribute the knowledge markup throughout the HTML as attributed of existing HTML tags

One page, not two

- Content providers prefer not to generate multiple pages, one for humans (HTML) and another for machines (RDF)
  - RDF serializations are complex
  - Requires a separate storage, generation, etc. mechanism
  - Introduces redundancy, which can lead to errors if we change one page but not the other
- Simplifies the job of search engines as well
### General approach

- Provide or reuse tag *attributes* to encode the metadata
  - Browsers and other web systems ignore attributes they don’t understand
- Three approaches have been developed
  - Microformats (~ 2005)
  - RDFa (~ 2007)
  - Microdata (~ 2012)

### Microformats approach

- Reuses HTML attributes like `@class`, `@title`
- Separate vocabularies (address, CV, ...)
- Difficult to mix microformats (no concept of namespaces)
- Does not, inherently, define an RDF representation
  - possible to transform via, e.g., XSLT + GRDDL, but all transformations are vocabulary dependent

### Microdata approach

- Defined and supported by Google and Bing
- Adds new attributes to HTML5 to express metadata
- works well for simpler “single-vocabulary” cases, but not well suited for mixing vocabularies or for complex vocabularies
- No notion of datatypes or namespaces
- Defines a generic mapping to RDF

### RDFa approach

- Adds new (X)HTML/XML attributes
- Has namespaces and URIs at its core
  - So mixing vocabulary is easy, as in RDF
- Complete flexibility for using literals or URI resources
- Is a complete serialization of RDF
Yielding this RDF:

```xml
<http://www.ivan-herman.net/foaf#me>
    schema:alumniOf <http://www.elte.hu> ;
    foaf:schoolHomePage <http://www.elte.hu> ;
    schema:worksFor <http://www.w3.org/W3C#data> ;

<http://www.elte.hu>
    dc:title "Eötvös Loránd University of Budapest" .

<http://www.w3.org/W3C#data>
```
Yielding this RDF

```
[ rdf:type schema:Review ;
  schema:name "Oscars 2012: The Artist, review" ;
  schema:description "The Artist, an utterly beguiling..." ;
  schema:ratingValue "5" ]
```
Rich Snippets

- Search engines add a few lines of text under results, giving users an idea of what's on the page and why it's relevant to their query
- These are often extracted from structured data embedded on the page

RDFa and microdata: similarities

- RDFa and Microdata are modern options
  - [Microformats](http://www.mwww.com/microformats/) is another
- Both have similar approaches
  - Structured data encoded in *HTML attributes only* – no new elements
  - Define some special *attributes*
    - e.g., `itemscope` for microdata, `resource` for RDFa
  - Reuse some HTML core attributes (e.g., `href`)
  - Use textual content of HTML source, if needed
- RDF data can be extracted from both

RDFa and microdata: differences

- Microdata *optimized* for simpler use cases:
  - One vocabulary at a time
  - Tree shaped data
  - No datatypes
- RDFa provides full serialization of RDF in XML or HTML
  - Price is extra complexity over Microdata
- RDFa 1.1 Lite is a simplified authoring profile of RDFa, very similar to microdata
Structured data in HTML is increasing

... 25% of webpages containing RDFa data [...] over 7% of web pages containing microdata.

Mail from Peter Mika, Yahoo!
Based on a crawl evaluation by P. Mika and T. Potter
LDOW2012 Workshop, April 2012, Lyon, France

... web pages that contain structured data has increased from 6% in 2010 to 12% in 2012.

Hannes Mühleisen and Christian Bizer
Web Data Commons—Extracting Structured Data from Two Large Web Corpora,
LDOW2012 Workshop, April 2012, Lyon, France