

Embedding Knowledge in HTML

Some content from a presentations by Ivan Herman of the W3c

Overview

- Why we want to embed structured data in HTML
- RDFa
- Microdata and schema.org
- RDFa lite as an encoding for Microdata
- JSON-LD as an encoding for RDF and Microdata
- Usecases and examples

HTML is Everywhere

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

Adding RDF-like data to HTML

- Like to add semi-structured knowledge in HTML
 - Humans see and understand regular HTML content (text, images, videos, audio)
 - Machines see and understand data markup in XML, RDF or some other format
- Possibilities include
 - Add a link to separate document with knowledge
 - Add new HTML tags
 - Embed knowledge as comments, Javascript, etc.
 - Distribute knowledge markup throughout HTML as attributes of existing HTML tags

One page, not two

- Content providers prefer not to generate multiple page:, one for humans (HTML) and another for machines (RDF)
 - RDF serializations are complex
 - Requires separate mechanisms for storage, generation, etc.
 - 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

- Two embedding techniques
 - Provide/reuse tag *attributes* to encode the metadata; browsers/apps ignore attributes they don't understand
 - Add a graph for page in JSON-LD
- Several approaches have been developed
 - [Microformats](#) (~ 2005)
 - [RDFa](#) (~ 2007)
 - [Microdata](#) (aka schema.org) (~ 2012)
- Status 2014/5 (IMHO)
 - *Microformats* still used but future is limited
 - *RDFa* and JSON-LD are preferred encoding of choice
 - Schema.org vocabularies getting large uptake



microformats

- Earliest idea, supplanted by RDFa and Microdata
- Reuses HTML attributes like @class, @title
- Separate vocabularies developed for common use cases, e.g., address, CV, recipes ...
- Difficult to mix microformats (no concept of namespaces)
- Doesn't define an RDF representation
 - possible to transform via, e.g., XSLT + GRDDL, but transformations are vocabulary dependent



microformats

- vCard: popular format for “business card” data
- Example use case for email
 - Sender attaches vCard to email message
 - Recipient detaches to contact app
- hCard is a Microformat based on vCard
 - Allows way to embed vCard data in a web page



microformats

```
BEGIN:VCARD
VERSION:4.0
N:Forrest;Gump;;Mr.;
FN:Forrest Gump
ORG:Bubba Gump Shrimp Co.
TITLE:Shrimp Man
PHOTO;MEDIATYPE=image/gif:http://www.example.com/dir_photos/my_photo.gif
TEL;TYPE=work,voice;VALUE=uri:tel:+1-111-555-1212
TEL;TYPE=home,voice;VALUE=uri:tel:+1-404-555-1212
ADR;TYPE=WORK,PREF;;;100 Waters Edge;Baytown;LA;30314;United States of
Amer
ica
LABEL;TYPE=WORK,PREF:100 Waters Edge\nBaytown\, LA 30314\nUnited States of
America
ADR;TYPE=HOME;;;42 Plantation St.;Baytown;LA;30314;United States of America
LABEL;TYPE=HOME:42 Plantation St.\nBaytown\, LA 30314\nUnited States of Ame
rica
EMAIL:forrestgump@example.com
REV:20080424T195243Z
END:VCARD
```



```
<ul class="vcard">
  <li class="fn">Forrest Gump</li>
  <li class="org">Bubba Gump Shrimp Co.</li>
  <li class="tel">1-111-555-1212</li>
  <li><a class="url" href="http://bubbagump.com/">
http://bubbagump.com/</a></li>
</ul>
```

- Recipient detaches to contact app
- hCard is a Microformat based on vCard
 - Allows way to embed vCard data to a web page

Microformats => Microdata

- Microformat problems:
 - Not integrated well into HTML
 - Not supported by a standardization body
 - Limited vocabulary
- Microformats have been supplanted by Microdata

Microdata approach

- Defined and supported by Google, Bing, Yahoo and Yandex
- Adds new attributes to HTML5 to express metadata
- Works well for simple “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



Ivan Herman

Who am I?

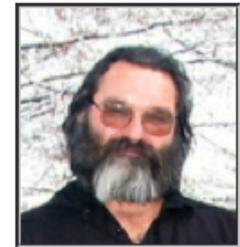
I graduated as mathematician at the [Eötvös Loránd University of Budapest](#), Hungary, in 1979. After a brief scholarship at the Université Paris VI I joined the Hungarian research institute in computer science ([SZTAKI](#)) where I worked for 6 years (and turned into a computer scientist...). I left Hungary in 1986 and, after a few years in industry in Munich, Germany, I joined the [Centre Mathematics and Computer Sciences \(CWI\)](#) in Amsterdam where I have a tenure position since 1988. I received a PhD degree in Computer Science in 1990 at the [University of Leiden](#), in the Netherlands. I joined the [World Wide Web Consortium \(W3C\)](#) Team as Head of [W3C Offices](#) in January 2001 while maintaining my position at CWI. I served as Head of Offices until June 2006, when I was asked to take the [Semantic Web Activity](#) Lead position, which is now my principal work at W3C.

Before joining W3C I worked in quite different areas (distributed and dataflow programming, language design, system programming), but I spend most of my research years in computer graphics and information visualization. I also participated in various graphics related ISO standardization activities and software developments. My ["professional" homepage](#) contains a list of [my publications](#) (see also [my Mendeley account](#)), [my public presentations](#), and details of the various projects I participated in the past. There is also a [dblp entry for my publications](#) generated automatically (although I am not sure it is complete...). (B.t.w., based on my publications, my [Erdős number](#) is ≤ 4 ...)

In my previous life (i.e., before joining W3C...) I was member of the Executive Committee of the [Eurographics Association](#) for 15 years, and I was vice-chair of the Association between 2000 and 2002. I was the co-chair of the [9th World Wide Web Conference](#), in Amsterdam, May 2000; since then, I have also been member of [IW3C2 \(International World Wide Web Conference Committee\)](#), responsible for the World Wide Web Conference series. Since autumn 2007 I am also member of [SWSA \(Semantic Web Science Association\)](#), the committee responsible for the International Semantic Web Conferences (better known as "ISWC") series.

Some personal data

- The Hungarian spelling of my [full name is Herman Iván](#). I.e, my name is Ivan (well, spelled properly: Iván) and my [surname is Herman](#) (many in the Netherlands and in Germany mix it up, and use "Herman" as my name... this is aggravated by the fact that, uniquely in Europe, the Hungarian custom is to put surname first).
- Nationalities: French and Hungarian
- Gender: male
- Family: I am married and have a son, David.
- Date and city of birth: 24th February, 1955, [Budapest](#), Hungary
- Email addresses: 'ivan' on my own ivan-herman.net domain, 'ivan' on the w3.org domain, or 'ivan.herman' on the cwi.nl domain
- (Mobile) Phone: +31-641044153
- Skype ID: ivan_herman
- I live in [Amstelveen](#) (see also [geonames](#)), the Netherlands (lat: 52.302063, long: 4.87397). This is a suburb of [Amsterdam](#). The closest airport is Amsterdam Schiphol
- I am the administrator of the [Semantic Web Activity Blog](#) at W3C which can either be [accessed directly](#) or via [its](#)



some links

- [personal homepage](#)
- more data on me
- [personal blog \(RSS feed\)](#)
- [homepage](#) at W3C
- ["professional" homepage](#)
- ["official" CV](#)
- [more about me](#)
- [my photos](#)

"social" links

- [facebook](#)
- [flickr](#)
- [picasa web](#)
- [linkedin](#)
- [IWIW](#)
- [tripit](#)
- [twitter](#)
- [Mendeley](#)
- [Google+](#)
- [freebase](#)


```

<meta property="foaf:accountName" content="ivan-herman" />
</a>
</li>
<li>
  <a href="https://plus.google.com/u/0/113268051484517627727" typeof="foaf:OnlineAccount">
    <span property="foaf:accountServiceHomepage" href="http://www.mendeley.com/">Google</span>
    <meta property="foaf:accountName" content="113268051484517627727" />
  </a>
</li>
<li>
  <a about="http://www.ivan-herman.net/foaf#me" rel="owl:sameAs" resource="http://rdf.freebase.com/ns/en.
</li>
</ul>
</div>
</div>
<div id="content" >
  <h1 property="schema:name foaf:name">Ivan Herman</h1>
  <meta property="foaf:title" content="Dr" />
  <h2>Who am I?</h2>
  <p>I graduated as mathematician at the <a rel="foaf:schoolHomepage" schema:alumniOf" href="http://www.elte.hu/"><span
I joined the <a rel="schema:worksFor" href="http://www.w3.org" resource="http://www.w3.org/Data#W3C">
  <span property="dc:title">World Wide Web Consortium (W3C)</span>
</a> Team as Head of <a rel="foaf:pastProject" href="http://www.w3.org/Consortium/Offices"><span property="dc:title"
<link rel="owl:sameAs" href="http://www.ivan-herman.net/me" />
<link rel="owl:sameAs" href="http://www.ivan-herman.net/Ivan_Herman" />
<link rel="foaf:workplaceHomepage" href="http://www.w3.org"/>
<meta property="schema:jobTitle" content="Semantic Web Activity Lead" />

<p>Before joining W3C I worked in quite different areas (distributed and dataflow programming, language design, syst
</p>

<p>In my previous life (i.e., before joining W3C...) I was member of the Executive Committee of the <a rel="foaf:pastI
</p>
<h2>
Some personal data
</h2>
<ul>
  <li>The Hungarian spelling of my full name is <span property="foaf:name" lang="hu">Herman Iván</span>.
  Ie, my name is <span property="foaf:givenname schema:givenName">Ivan</span> (well, spelled properly:
  <span property="foaf:givenname schema:givenName" lang="hu">Iván</span>) and my surname is
  <span property="foaf:surname schema:familyName">Herman</span>
  (many in the Netherlands and in Germany mix it up, and use "Herman" as my name... this is aggravated by the fact t
</li>
  <li>Nationalities: <span property="schema:nationality">French</span> and <span property="schema:nationality">W

```

Yielding this RDF

```
<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>
  dc:title "World Wide Web Consortium (W3C)"
...
```

The Telegraph

Search - enhanced by Google

Monday 09 April 2012

HOME NEWS SPORT FINANCE COMMENT BLOGS **CULTURE** TRAVEL LIFESTYLE FASHION TECH Dating Offers Jobs

Film Music Art Books TV and Radio Theatre Hay Festival Dance Opera Photography Comedy Video In the Know

Oscars Film Reviews Cinema Trailers Coming Soon Talking Movies Interviews DVDs Film Life Film Video

HOME » CULTURE » FILM » FILM REVIEWS

Oscars 2012: The Artist, review

The Artist, an utterly beguiling silent, black-and-white celebration of early Hollywood won Best Picture at the Oscars 2012.

★★★★★



To attend an Info Session or Masterclass Or download a brochure

[Click here](#)

INSEAD
The Business School for the World*

Global Executive MBA

+1

TELEGRAPH TICKETS »



The Telegraph

Search - enhanced by Google

Monday 09 April 2012

- HOME
- NEWS
- FINANCE
- COMMENT
- BLOGS
- CULTURE**
- TRAVEL
- LIFESTYLE
- FASHION
- TECH
- Dating
- Offers
- Jobs

- TV
- Radio
- TV and Radio
- Theatre
- Big Festival
- Dance
- Opera
- Photography
- Comedy
- Video
- In the Know

- Books
- Games
- Comic Strips
- Country Songs
- Talking Movies
- Interviews
- DVDs
- Film Life
- Film Video

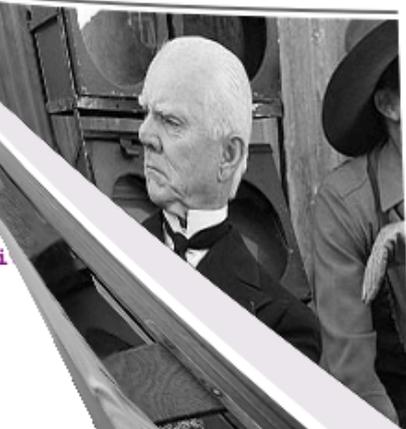
```

<li class="...">
</li><a href="...">
</div>
</div>
<!-- googleon: all -->
<div id="tmglBody" >
  <div class="access"><a name="a...">
  </div>
  <div class="twoThirdsThird2 gutter">
    <div class="twoThirds gutter">
      <div class="story">
        <h1 itemprop="name">Oscars 2012: The Artist, review
        <h2 itemprop="description">
The Artist, an utterly beguiling silent, black and white celebration of early Hollywood won Best Picture at the Oscars 2012
        </h2>
        <div class="rating" itemprop="reviewRating">
          <meta itemprop="worstRating" value="1">
          <meta itemprop="bestRating" value="5">
          <span itemprop="ratingValue" value="4">
            
          </div>
          <div class="artIntro">
            <div id="storyEm...">
              <div class="slideshow ssIntro">
                <div class="nextPrevLayer">
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</div>
<!-- remove the whitespace added by escenic before end of

```

Oscars 2012: The Artist, review

The Artist, an utterly beguiling silent, black and white celebration of early Hollywood won Best Picture at the Oscars 2012



To attend an Info Session or Masterclass Or download a brochure

[Click here](#)

INSEAD
The Business School for the World®

Global Executive MBA

TELEGRAPH TICKETS »



```
<li class="first"><a href="/">Home</a><span>&raquo;</span></li>
<li><a href="http://www.telegraph.co.uk/culture/">Culture</a><span>&raquo;</span></li>
  <li><a href="http://www.telegraph.co.uk/culture/film/">Film</a><span>&raquo;</span></li>
  <li class="styleSix"><a href="http://www.telegraph.co.uk/culture/film/filmreviews/">Film R
</div>
</div>
<!-- googleon: all -->
<div id="tmglBody" >
  <div class="access"><a name="article"></a></div>

  <div class="twoThirdsThird2 gutterUnder">
    <div class="twoThirds gutter" itemscope itemtype="http://schema.org/Review">
      <div class="storyHead">
        <h1 itemprop="name">Oscars 2012: The Artist, review</h1>
        <h2 itemprop="description">
The Artist, an utterly beguiling silent, black-and-white celebration of early
Hollywood won Best Picture at the Oscars 2012.
</h2>
        <div class="rating" itemprop="reviewRating" itemscope itemtype="http://schema.org/Rating">
          <meta itemprop="worstRating" content = "0.5">
          <meta itemprop="bestRating" content = "5">
          <span itemprop="ratingValue" class="hidden">5</span>
          
        </div>
        <div class="artIntro">
          <div id="storyEmbSlide">
            <div class="slideshow ssIntro">
              <div class="nextPrevLayer">
                <div class="ssImg">
                  
                    <div class="ingCaptionCredit">
                      <span class="caption">B er enice Bejo as ris
                    </div>
                  </div>
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
<div class="oneHalf gutter">
  <div class="story">
    <div class="cl"> </div>
    <!-- remove the whitespace added by escenic before end of </a> tag -->
```

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 text under results to preview what's on page and why it's relevant
- Text often extracted from structured data embedded on the page
- See <http://bit.ly/RichSN> for more information

[Little Water Cantina - Eastlake - Seattle, WA](#)
www.yelp.com › Restaurants › Mexican
★★★★☆ 90 reviews - Price range: \$\$
90 Reviews of **Little Water Cantina**. "Three things are on my list when I eat out: great food, atmosphere, and [Vegetarian Vegan Pizza No Cheese\) Recipe - Food.com - 248865](#)"

 [www.food.com/recipe/vegetarian-vegan-pizza-no-c...](#)
★★★★★ 2 reviews - 1 hr 32 mins - 242.9 cal
Aug 26, 2007 – This is from my dad, who developed some **vegan recipes** doesn't have any cheese, and you

[Leonard Cohen – Free listening, videos, concerts, stats, & pictures at...](#)
www.last.fm/music/Leonard+Cohen
Watch videos & listen to **Leonard Cohen**: Suzanne, Hallelujah & more, plus 132 pictures. **Leonard Cohen**, (born September 21, 1934 in Montréal, Quebec, ...

Track	Duration
Suzanne	♫ 3:48
The Darkness	♫ 4:29
Going Home	♫ 3:51
Hallelujah	♫ 6:12

Everything

Images

Maps

Videos

News

Shopping

More

Amsterdam

Change location

Any time

Past hour

Past 24 hours

Past week

Past month

Past year

Custom range...

More search tools

The Artist showtimes for Amsterdam

[Pathe Tuschinski](#) - Reguliersbreestraat 26-34, Amsterdam - [Map](#)
11:50 - 14:05 - 19:10

[Filmtheater "De Uitkijk"](#) - Prinsengracht 452, Amsterdam - [Map](#)
12:15 - 19:00 - 21:15

[Filmtheater Rialto](#) - Ceintuurbaan 338, Amsterdam - [Map](#)
12:45

[+ Show more theaters](#)

The Artist (2011) - IMDb

www.imdb.com/title/tt1655442/

Silent **movie** star George Valentin bemoans the coming era of talking ... Still of Jean Dujardin and Missi Pyle in **The Artist** Still of Bérénice Bejo in **The Artist** Reem ...

[Full cast and crew](#) - [The Artist Trailer \(Official ...](#) - [Bérénice Bejo](#) - [Jean Dujardin](#)

The Artist (film) - Wikipedia, the free encyclopedia

[en.wikipedia.org/wiki/The_Artist_\(film\)](http://en.wikipedia.org/wiki/The_Artist_(film))

The Artist is a 2011 French romantic comedy drama in the style of a black-and-white silent **film** written and directed by Michel Hazanavicius, starring Jean ...

[Jean Dujardin](#) - [Bérénice Bejo](#) - [Uggie](#) - [Diegesis](#)

The Artist Trailer 2011 HD - YouTube



www.youtube.com/watch?v=O8K9AZcSQJE

25 Aug 2011 - 3 min - Uploaded by TrailersApplecom
I love how George Clooney, and Brad Pitt, lost the Best actor category to this **film**. It just shows that there is ...

[More videos for the artist movie »](#)

Oscars 2012: The Artist, review - Telegraph

www.telegraph.co.uk/Culture/Film/Film_Reviews

★★★★★ Review by Robbie Collin

27 Feb 2012 - **The Artist**, the final **film** to be released in 2011 and also the most heart-sweetly joyful one, is a silent **movie**, screened in black and white and ...

[The Artist is the perfect film about Hollywood | Hadley Freeman](#)

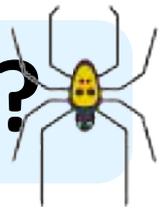
RDFa and Microdata: similarities

- RDFa and Microdata are modern options
- 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

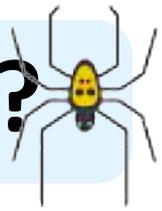
- Microdata *optimized* for simpler use cases:
 - One vocabulary at a time
 - Tree shaped data
 - Few datatypes
- RDFa provides full serialization of RDF in XML or HTML
 - Price is extra complexity over Microdata
- RDFa 1.1 Lite is a simplified authoring profile of RDFa, very similar to microdata

Amount of structured data on Web?



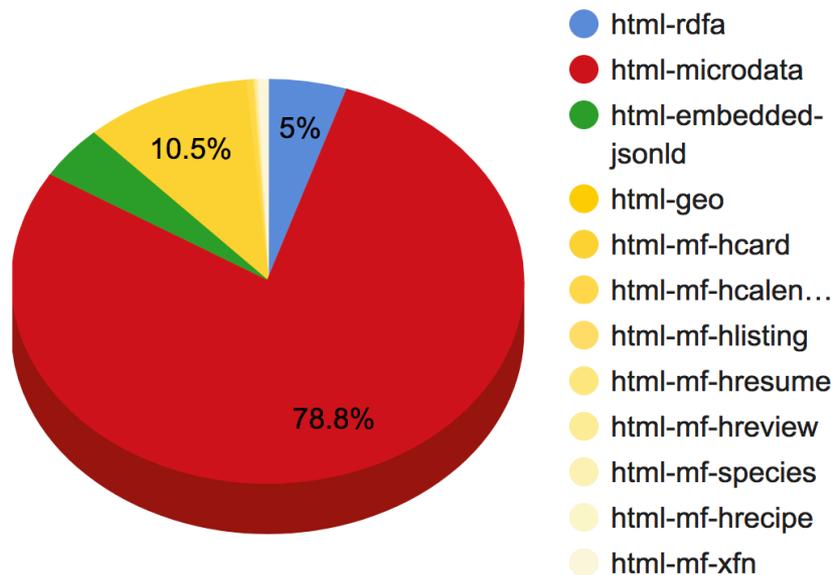
- Web Data Commons project uses Common Crawl data to estimate amount of structured data on Web
- Looks for Microdata, RDFa, other formats (e.g., hCalendar, hCard) in URLs parsable as HTML
- Oct. 2016 crawl analyzed 3.2B HTML pages:
 - 1.2B pages (30%) with structured data in 5.6M domains of 34M (17%)
 - 44B triples about 9B typed entities
- Data can be downloaded

Amount of structured data on Web?



Top Domains by Extracted Triples

Triples



1. blogspot.com (613,297,268 triples)
2. ticketprocess.com (470,790,528 triples)
3. moosejaw.com (410,021,577 triples)
4. hallmark.com (351,800,787 triples)
5. theclothdiaperwhisperer.com (314,888,334 triples)
6. cnbc.com (298,241,469 triples)
7. hotels.com (288,853,467 triples)
8. repairpal.com (266,250,224 triples)
9. uncommongoods.com (261,956,686 triples)
10. wordpress.com (219,485,949 triples)
11. justia.com (207,590,139 triples)
12. leadferret.com (207,038,617 triples)
13. propartner.ru (180,976,524 triples)
14. callersmart.com (162,779,369 triples)
15. gigmasters.com (151,575,208 triples)
16. epicsports.com (150,282,774 triples)
17. unitiki.com (125,549,640 triples)
18. drom.ru (122,572,723 triples)
19. zap2it.com (122,118,800 triples)
20. caasa.it (108,318,384 triples)
21. [More](#)



FREE SHIPPING on Orders \$35+

What are you looking for?



Brands

Jackets

Clothing

Footwear

Hike & Camp

Climb

Snow

More Activities

Sale

Free Almost Gift Card up to \$100 - Use Code: FREEMONEY | Free 2-Day Shipping on Jackets over \$99 [DETAILS](#)

Home | Womens Jackets | Womens Canada Goose Jackets



VIDEO



Canada Goose Women's Shelburne Parka

★★★★★ 4.8

\$925.00

Select a Size



[Size Guide](#)

Select a Color



Qty:

1

This Item SHIPS FREE

When will I get it?

Add To Cart

Call 877-MOOSEJAW | [Gift Cards](#) | [Store Locator](#)

[OUTDOOR KNOWLEDGE](#) | [MADNESS](#)

Sign-up for Our Email

GO



Microdata RDFa POSH

Statement Collection #1

Entity https://www.moosejaw.com/moosejaw/shop/product_Canada-Goose-Women-s-Shelburne-Parka_10254734_10208_1000001_-1

Attributes

rdf:type	schema:Product
schema:offers	See Statement Collection #2
schema:offers	See Statement Collection #4
schema:offers	See Statement Collection #6
schema:offers	See Statement Collection #8
schema:offers	See Statement Collection #10
schema:offers	See Statement Collection #12
schema:offers	See Statement Collection #14
schema:offers	See Statement Collection #16
schema:offers	See Statement Collection #18
schema:offers	See Statement Collection #20
schema:offers	See Statement Collection #22
schema:offers	See Statement Collection #24
schema:offers	See Statement Collection #26
schema:offers	See Statement Collection #28
schema:offers	See Statement Collection #30
schema:offers	See Statement Collection #32
schema:offers	See Statement Collection #34
schema:offers	See Statement Collection #36
schema:offers	See Statement Collection #38
Name	Canada Goose Women's Shelburne Parka
schema:brand	Canada Goose
schema:image	//s7d1.scene7.com/is/image/MoosejawMB/10254734x1018148_zm?product1000\$

Navigation: & Camp, Climb, Snow, More Activities, Sale

Red Banner: **NEW** | Free 2-Day Shipping on Jackets over \$99 | DETAILS

Canada Goose Women's Shelburne Parka

★★★★★ 4.8

\$925.00

Select a Size

[Size Guide](#)

Select a Color

Qty: **This Item SHIPS FREE**
When will I get it?

Add To Cart

Conclusions

- The amount of structured data on the web is growing steadily
- Microdata shows the strongest growth
- RDFa also common
- Microformat data is probably not growing as much