Kit Sen Chin, 123RF.com

Working with microformats and microdata

Hidden Meaning

Programs aren't as smart as humans when it comes to interpreting the meaning of web information. If you want to maximize your search rank, you might want to dress up your HTML documents with microformats and microdata. *By Andreas Möller*

HTML lets you mark up sections of text as headings, body text, hyperlinks, and other web page elements. However, these definitions have nothing to do with the meaning of the data: Does the text refer to a person, an organization, a product, or an event? Microformats [1] and their successor, microdata [2] make the meaning a bit more clear by pointing to business cards, product descriptions, offers, and event data in a machine-readable way.

In this article, I describe some important microformats, such as hCard, hCal, hProduct, hReview, and Geo. You'll also get an introduction to microdata, and you'll learn about some open source tools for integrating awareness of microformats and microdata into your own programs.

Microformats

HTML was originally designed for humans to read, but with the explosive growth of the web, programs such as search engines also process HTML data. What do programs that read HTML data typically find? Listing 1 shows the HTML source code for the website shown in Figure 1 – an HTML5 document with a business card. The Heading text block is marked up with the element `h1`. The text for the business card is surrounded by the `div` container element, and `<br/>` introduces a line break.

It is easy for a human reader to see that the data shown in Figure 1 represents a business card – even if the text is

LISTING 1: HTML5 Document with Business Card

```
01 <!DOCTYPE html>
02 <html lang="es">
03 <head></head>
04 <body>
05  <h1>Pedro Miguel Díaz - el Arte Guitarra Flamenca</h1>
06  <div>
07   <img alt="" src="images/el-fumador.jpg"/>
08   <a href="http://el-fumador.info">Pedro Miguel Díaz</a> - el Fumador<br/>
09   Flamenco Artist<br/>
10   C/ Peral 43<br/>
11   E-41002 Sevilla<br/>
12   España<br/>
13   pedro@fumador.info<br/>
14   +34 954 88 24 37
15  </div>
16 </body>
17 </html>
```

AUTHOR

Andreas Möller, Dipl. Phys. [*http://pamoller.com*], has focused on developing Internet-based software for more than 10 years. His development work includes database applications, web applications, and software in the field of single-source publishing. He currently works as a freelance consultant and author.

in a different language, but a computer program would not know the difference between this data and any other data enclosed in HTML tags. Microformats provide a means of showing the purpose of the data. Listing 2 shows the microformat markup for a typical business card. As you can see, microformats mainly use the `class` attribute from the HTML vocabulary. In this case, the hCard microformat describes a business card.

In Listing 2, every property is enclosed by the `span` element and described by the universal `class` attribute. For example, `class="vcard"` in line 6 refers to a business card object. The `class="fn url"` assignment describes the contents of the hyperlink, `Pedro Miguel Díaz`, as a name (`fn`, "Full name") and the URL in the `href` attribute as the matching website (`url`).

The address object `class="adr"` in line 11 comprises multiple properties. It contains the `type`, `street-address`, `postal-code`, `locality`, and `country` of the address. The business card's markup follows the hCard [3] microformat, which implements the Internet Mail Consortium's vCard specification [4] in HTML. For each line of text, vCard stores the key/value pair and additional options. The hCard microformat translates this vocabulary into HTML by translating the keys to values of the `class` attribute.

Microformats exist for several other uses in addition to business cards. For example, you can describe events using hCalendar, geographical information with Geo, and products and product reviews with hReview and hReview-aggregate. For an overview of microformats and versions, check out the wiki at Microformats.org [5].

Figure 1: A website with the most important data for a musician – hosting service providers also refer to this as a web business card.

LISTING 3: Event Markup with hCalendar

```
01 <div class="vevent">
02  <h3 class="summary">Feria de Abril</h3>
03  <abbr class="dtstart" title="2012-03-23T21:00:00Z">el 23 de marzo de 2012,
       21:00</abbr>
04  en <span class="location">Sevilla</span>
05  <div class="geo">
06   <abbr class="latitude" title="37.3844">37° 38&#8242; N</abbr>
07   <abbr class="longitude" title="-5.9888">5° 98&#8242; W</abbr>
08  </div>
09 </div>
```

Listing 3 shows how to use the hCalendar microformat, which was derived from iCalendar [6]. The property/value pair `class="vevent"` describes the event, which must contain a `summary` and a start (`dtstart`). In contrast, the `location` information is optional.

The following rules apply for microformats: if you want to provide a date in a machine-readable format, it has to be given in the `title` property of the `abbr` element (line 3).

Besides grammar rules, microformats are characterized by their reusability. Listing 3 uses the Geo microformat to specify the location for the event. The format contains properties for the `latitude` and `longitude`. Again, the machine-readable variant is given in the title property of the `abbr` element.

Listing 4 provides another example of the re-usability of microformats. Line 1 marks of the summary of a product review using the hReview-aggregate micro-

LISTING 2: Business Card with hCard Markup

```
01 <!DOCTYPE html>
02 <html>
03 <head><link rel="profile"
      href="http://microformats.org/profile/hcard"/></head>
04 <body>
05  <h1>Pedro Miguel Díaz - el Arte Guitarra Flamenca</h1>
06  <div id="me" class="vcard">
07   <img alt="" class="photo" src="images/el-fumador.jpg"/>
08   <a class="url fn" href="http://el-fumador.info">Pedro
       Miguel Díaz</a>
09   - <span class="nickname">el Fumador</span><br/>
10   <span class="title">Flamenco Artist</span>
11   <div class="adr">
12    <span style="display:None" class="type">Home</span>
13    <span class="street-address">C/ Peral 43</span><br/>
14    <span class="postal-code">E-41002</span>
15    <span class="locality">Sevilla</span><br/>
16    <span class="country-name">España</span>
17   </div>
18   <span class="email">pedro@fumador.info</span><br/>
19   <div class="tel">
20    <span class="type" style="display:None">Home</span>
21    <span class="value">+34 954 88 24 37</span>
22   </div>
23  </div>
24 </body>
25 </html>
```

LISTING 4: Product Markup with hProduct

```
01 <li class="hreview-aggregate">
02  <span class="item">
03   <span class="hproduct">
04    <a class="fn url" href="www.reseller.
        com?isbn=9-874-444-333-85-1">El Arte Flamenco</a>
05    <span class="price">33.90&#8364;</span>
06    <span class="identifier">
07     <span class="type">ISBN</span>
08     <span class="value">9-874-444-333-85-1</span>
09    </span>
10   </span>
11  </span>
12  <abbr class="rating" title="4.9"><sup>&#x2606;&#x2606;
      &#x2606;&#x2606;&#x2606;</sup></abbr>
13  <span class="votes">200</span>
14 </li>
```

format. hProduct is used to state the product as an `item`. The `rating` refers to the overall score, and `votes` to the number of reviews.

Microformats use the `rel` attribute in hyperlinks to express the relationship to the target resource; for example, the link:

```
<a rel="met friend colleague" ➥
 href="http://esoleares.es">➥
 Estrella Soleares - la Pinta</a>
```

refers to the homepage belonging to Estrella Soleares. The list in the `rel` attribute designates Estrella as a friend and colleague who the link provider has also (`met`) personally. For an overview of values for the `rel` attribute, again check out the Microformats wiki [7].

HTML5 Microdata

HTML5 includes a number of semantic elements that weren't present in early HTML versions. The `time` element is used to mark up dates and times; the `meter` element signifies measured values. To make this semantic data more understandable for programs, HTML5 adapts the microformat concept into the form known as microdata. The HTML5 microdata system defines a set of properties – `itemscope`, `itemtype`, `itemprop`, and `itemref` – that provides the function of the `class` attribute in microformats.

Listing 5 shows the business card data from Listing 2 as HTML5 microdata. Instead of a `class` property, microdata uses the `itemprop` attribute to express properties. The `itemscope` attribute is used to delimit units, as in lines 10 through 17.

The `itemtype` attribute (line 5) designates the vocabulary for the section it introduces. The HTML5 standard has adopted the vocabulary for vCard and vEvent. Other microformats have also been translated into microdata [8].

The `itemref` attribute is used to group properties beyond container logic. To allow this to happen, `itemref` uses the universal ID attribute to point to other elements.

Rich Snippet Tool

The significance of microformats doesn't become apparent until you actually view the data prepared with them. For example, you can use microformats to optimize the results obtained by search engines. Figure 2 shows the effect of the hCard microformats on the Google search engine using the Rich Snippet Tool [10]. The Rich Snippet Tool picks up the microformats from the page and shows which information the search engine evaluates. For the website shown in Figure 1, the search engine picks up Pedro Miguel's location (Sevilla) and the job designation (Flamenco Artist). Reviews based on hReview or hReview-aggregate are also highlighted in the search results. The grades from 1 through 5 are encoded graphically as the number of stars.

It is difficult to estimate how widespread the more recent microdata has become compared with microformats;

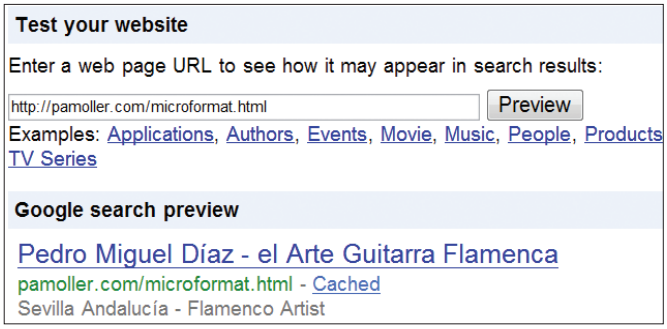

Figure 2: Google's Rich Snippet Tool shows which information a search engine picks up from the microformats and microdata.

LISTING 5: Business Card with Microdata

```
01 <!DOCTYPE html>
02 <html>
03 <head><title>vCard</title></head>
04 <body>
05  <div id="me" itemscope
      itemtype="http://microformats.org/profile/hcard">
06   <img alt=""
      itemprop="photo" src="images/el-fumador.jpg"/>
07   <a rel="me" itemprop="url fn"
      href="http://el-fumador.info">Pedro Miguel Díaz</a>
08   - <span itemprop="nickname">el Fumador</span><br/>
09   <span itemprop="title">Flamenco Artist</span>
10   <div itemscope itemprop="adr">
11    <span style="display:None" itemprop="type">Home</span>
12    <span itemprop="street-address">C/ Peral 43</span><br/>
13    <span itemprop="postal-code">E-41002</span>
14    <span itemprop="locality">Sevilla</span><br/>
15    <span itemprop="region">Andalucía</span><br/>
16    <span itemprop="country-name">España</span>
17   </div>
18   <span itemprop="email">pedro@fumador.info</span><br/>
19   <div class="tel" itemscope>
20    <span itemprop="type" style="display:None">Home</span>
21    <span itemprop="value">+34 954 88 24 37</span>
22   </div>
23  </div>
24 </body>
25 </html>
```

however, the number of open source tools available for this task does suggest increasing use. As the Rich Snippet Tool shows, Google reads both microdata and microformats.

Working with Microformats

Tools for microformats or microdata fall into two classes: One class creates and marks up the data, and the other class reads the information. For example, the free content management system, WordPress, has an extension that facilitates the process of creating microdata [11].

The Operator plug-in [12] for the Firefox browser simplifies the process of identifying and processing microformats on websites. After installing the add-on and restarting the browser, Operator pops up as an additional toolbar below the address bar (Figure 3).

The current 0.9.5.6 version of Operator shows you Contacts, Events, Locations, Tagspaces, Bookmarks, and Resources and lets you process the data. Contacts can be exported to the vCard format and Events to iCal format, or they can be injected into web services such as Yahoo Contacts or Google Calendar. The plugin will pass Geo coordinates into map services if desired. Operator is released under the MPL/GPL/LGPL license. Future versions promise more microformats and links.

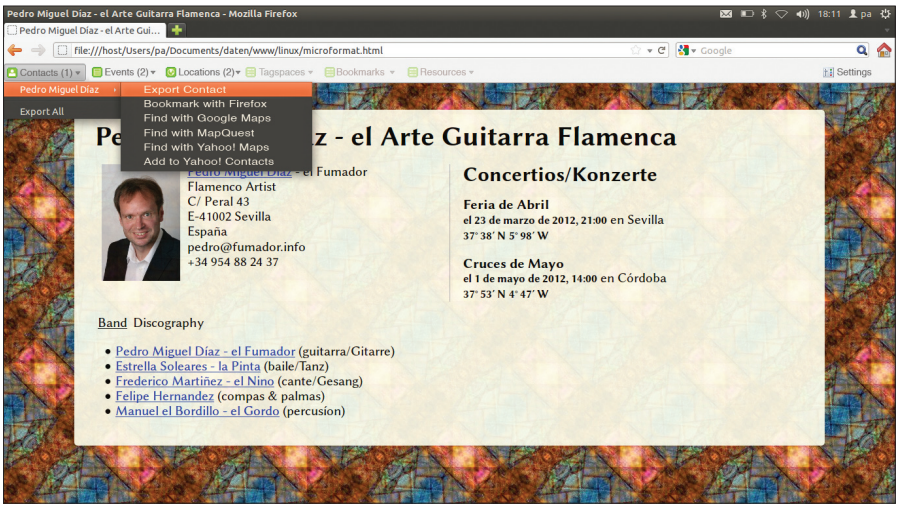

Figure 3: The Firefox Operator add-on gives users convenient access to information from microformats.

A number of open source libraries in various languages provide programmatic access to microdata. The Jquery plugin MicrodataJS [13] offers a programming interface similar to the DOM API [14] for microdata from the HTML5 specification. Listing 6 shows the integration of MicrodataJS in an HTML document. Line 3 integrates Jquery, then lines 4 and 5 integrate the two Jquery plugins, MicrodataJS and Microdata vCard.

Listing 7 demonstrates the use of MicrodataJS. This example uses JavaScript to extract the microdata for business cards and highlights them in vCard format. After loading the HTML document, the code in Listing 7 executes the function assigned to the `onload` event in line 1. The `$` function in line 2 expects a CSS selector expression. It looks for elements that contain `itemtype` attributes with the specification of the hCard vocabulary.

`each()` passes the resulting set of elements to the subsequent function. The `item` variable in line 3 accepts a reference to the current element from the special `this` variable. This value is then visible in all expressions, including the function definition that starts in line 6. If the user then executes the function in line 6 by clicking a hyperlink with the `class="microlink"` property, JavaScript remembers the value of `item`.

Before this can happen, line 4 uses the `append` function to add a hyperlink with the property `class="microlink"` to every element selected in line 2. Line 5 assigns the function defined in line 6 to this hyperlink's click event. The `$(this).parent()` expression in line 7 also points to the value of the `item` variable.

It is easier to use the `item` variable, as shown in this code. After a click, the

LISTING 6: Integration of MicrodataJS

```
01 <head>
02  <script src="js/fumador.js"></script>
03  <script src="js/jquery-1.7.1.min.js"></script>
04  <script src="js/jquery.microdata.js"></script>
05  <script src="js/jquery.microdata.vcard.js"></script>
06 </head>
```

LISTING 7: MicrodataJS

```
01 window.onload = function() {
02  $('[itemtype="http://microformats.org/profile/hcard"]').
      each(function(index) {
03   var item = $(this);
04   item.append('<a class="microlink">microdata</a>');
05   item.find("a.microlink").click(
06    function() {
07     $(this).parent().prepend('<div class="microdata">
       <a class="microlink close">close</a><textarea >'
       +jQuery.microdata.vcard(item)+'</textarea></div>');
08     item.find('.close').click(
09      function() {
10       item.find('.microdata').remove()
11      }
12     );
13    }
14   );
15  });
16 }
```

LISTING 8: Installing Python Microdata

```
01 wget http://pypi.python.org/packages/source/m/microdata/
   microdata-0.3.0.tar.gz
02 tar xzvf microdata-0.3.0.tar.gz
03 cd microdata-0.3.0/
04 sudo python setup.py install
```

REAL SOLUTIONS FOR REAL NETWORKS

FREE CD or DVD in Every Issue!

Each issue delivers technical solutions to the real-world problems you face every day.

Learn the latest techniques for better:

- network security
- system management
- troubleshooting
- performance tuning
- virtualization
- cloud computing

on Windows, Linux, Solaris, and popular varieties of Unix.

Now 6 issues per year!

ORDER ONLINE AT: shop.linuxnewmedia.com

`prepend` function in line 7 adds an important area populated with data in vCard format to the `item` element. For this to happen, the `vcard` method from MicrodataJS is used – also in line 7. Listing 7 also defines an event handler for the `close` class hyperlink. Another click closes the area that was opened. Figure 4 shows the extracted microdata.

MicrodataJS can also output microdata that it reads as JSON, Plugin Microdata JSON, Vevent, or Plugin Microdata Vevent format. You will find a plethora of functions for working with microdata, which are – unfortunately – not documented anywhere outside of the source code. On top of this, MicrodataJS provides a front end called Live for extracting microdata in the browser.

The Python programming language offers the Microdata [15] library, which provides a parser for microdata in HTML5 documents. It is released as public domain without any specific details on the license. Listing 8 shows how to install Microdata 0.3.0 on Ubuntu 11.10. If needed, the Python `html5lib` is also installed in this process.

Listing 9 demonstrates the use of Microdata in an interactive Python session. Python is first called at the command line in line 1, and the Microdata module is imported in line 2. Calling the `get_items()` method against the open file handle of `microdata.html` in line 3 extracts all the microdata from the HTML5 document and stores the information in the `items` field. A call to the `json()` function in line 4 converts the microdata to the JSON data format. In a similar way, the `dict()` method could translate the data to a Python dictionary format.

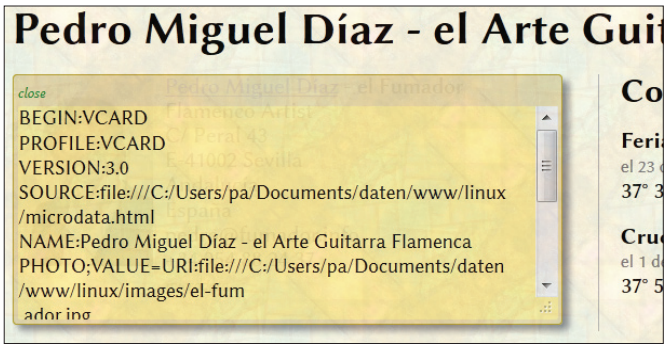

Figure 4: JavaScript and the MicrodataJS library put together this business card from microdata.

Other languages also offer free parsers for microdata; for example, Ruby in the form of Mida [16], Perl with HTML::Microdata [17], PHP with PHP Microdata [18], or Java with Any23 [19]. Any23 doesn't just parse microdata but converts a variety of meta-information between various formats. The supported input formats are: RDF, microdata, or microformats in (X)HTML5, Turtle, N-Triples, N-Quads, RDF in XML, and CSV. The supported output formats are Turtle, N-Triples, N-Quads, RDF/XML, and JSON.

LISTING 9: Microdata Interactive

```
01 $ python
02 >>> import microdata
03 >>> items = microdata.get_items(open("microdata.html"))
04 >>> print items[1].json()
05 {
06   "geo": [
07     {
08       "latitude": [
09         "37\u00b0 38\u2032 N"
10       ],
11       "longitude": [
12         "5\u00b0 98\u2032 W"
13       ]
14     }
15   ],
16   "dtstart": [
17     "2012-03-23T21:00:00Z"
18   ],
19   "type": "http://www.data-vocabulary.org/Event",
20   "location": [
21     "Sevilla"
22   ],
23   "summary": [
24     "Feria de Abril"
25   ]
26 }
```

Treasure Trove of Information

Microformats open up the content of web pages to smarter programmatic processing by tools such as search engines. HTML5 has adopted the microformat principle in the form of microdata. Just experimenting with the Firefox Operator plugin gives some idea of the potential these data formats offer. As more website providers adopt microformats and microdata, it will become easier to extract useful information from the ever more complex jumble of information we know as the web. ■■■

INFO

[1] Microformats: *http://microformats.org*
[2] Microdata: *http://www.whatwg.org/specs/web-apps/current-work/multipage/microdata.html*
[3] hCard: *http://microformats.org/wiki/hcard*
[4] vCard: *http://www.ietf.org/rfc/rfc2426.txt*
[5] Microformats wiki: *http://microformats.org/wiki/Main_Page*
[6] iCalendar: *http://www.ietf.org/rfc/rfc2445.txt*
[7] `rel` attribute: *http://microformats.org/wiki/existing-rel-values#HTML5_link_type_extensions*
[8] Schema.org: *http://schema.org*
[9] W3C Markup Validation Service: *http://validator.w3.org*
[10] Google Rich Snippet Tool: *http://www.google.com/webmasters/tools/richsnippets*
[11] Wordpress plugins: *http://wordpress.org/extend/plugins/microdata-for-seo-by-optimum7com/screenshots/*
[12] Firefox Operator add-on: *https://addons.mozilla.org/en-US/firefox/addon/operator/*
[13] MicrodataJS: *https://gitorious.org/microdatajs/*
[14] Microdata DOM API: *http://www.whatwg.org/specs/web-apps/current-work/multipage/links.html#microdata-dom-api*
[15] Python microdata extension: *http://pypi.python.org/pypi/microdata/0.3.0*
[16] Mida: *http://lawrencewoodman.github.com/mida/*
[17] HTML::Microdata: *http://search.cpan.org/~satoh/HTML-Microdata-0.02/lib/HTML/Microdata.pm*
[18] PHP Microdata: *https://github.com/soyrex/PHP-Microdata*
[19] Any23: *http://developers.any23.org*