ISSN No. 0976-5697

Volume 3, No. 7, Nov-Dec 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Semantic Annotation and Information Retrieval using Markup

Meena Unni*
Research Scholar, Department of Computer Science
Karpagam University, Coimbatore, India
meena_ukh@yahoo.com

Dr. K. Baskaran
Associate Professor, Department of CSE
Government College of Technology, Coimbatore
baski_101@yahoo.com

Abstract: Vision of the Semantic Web is to add meaning into the current web so machines can understand its contents. For the semantic web to succeed, it needs to take advantage of the existing web. New technologies have emerged to bridge the gap between current web and semantic web: microformats, RDFa and microdata. These technologies allow web page authors to embed extra semantic information within XHTML to mark up the structure. So when a given application accesses this semantic content, it will be able to tell more about the web page. Such a web page is readable by both human and machine. This paper describes these technologies with appropriate examples. It also explores their strength and weaknesses. Finally future considerations are illustrated.

Keywords: RDFa, microformat, microdata, smantic markup, annotation

I. INTRODUCTION

The Semantic Web [1] allows the representation and exchange of information in a meaningful way, facilitating automated processing of descriptions on the Web. A large number of technologies and software packages have emerged that allow interpretation of semantic information. This means to make the data available in RDF. One method is to encode RDF data in one of the serialization formats, like RDF/XML or turtle. This process makes publishers to make datasets available in specific formats. Interfaces are also being made to convert existing data into RDF often via SPARQL endpoint. Multiple approaches should be recognized to express web content in RDF depending on the type of data.

One of the major sources of data on the Web is HTML and XHTML. Web users realize that it is beneficial to reveal structured information in a web page than that can be displayed by a browser only for human consumption.

As of now retrieving data automatically involves scraping, which involves site specific APIs. This approach is efficient for few web sites. For scrapping, prior knowledge of the target page layout is required. If the page layout changes, then the screen scraper needs to be reconfigured. This resulted in XHTML pages including additional information to provide structured data.

This additional information is embeddeed to XHTML content as elements or attributes. For example, if an HTML element contains the words “Meena Unni”, then an extra attribute can denote the fact that this is the name of a person. More information can be given to represent the fact that Meena is the given name and Unni is the family name. Processors reading this XHTML can infer that the given name is the name of a person and process it accordingly.

Primary goal of semantic web is to search and automatically extract and process vast amount of information available on the web by machines. This paper focus on semantic markup technologies, which is used to include semantic Information (=metadata) in the source code that is the XHTML code. This process is called annotation, and it adds information to existing content which can be read by machines. Semantic Annotation enhances data with a context that is linked to structured knowledge.

Different technologies like microformats, RDFa and microdata are available to embed the meaning of the data into web pages. The greatest advantage of using these technologies is that it has no impact on the look and feel of the document.

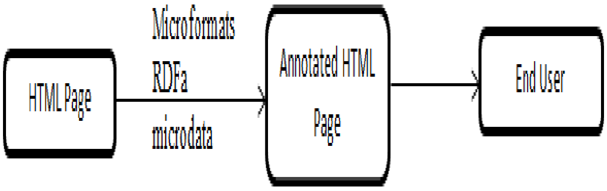


Figure 1: HTML page enriched with semantic markup

This paper after giving the introduction elucidates these technologies along with their strength and weaknesses. Finally application of these technologies is illustrated with an example. The paper culminates with conclusion and future work.

II. MICROFORMATS, RDFA AND MICRODATA

A. *Microformats:*

One of the technologies used to add information to web page is microformats. Microformats do not require any new standards. HTML tags are reused as much as possible. Web pages are made semantically rich by using class attributes with standardized properties. This ensures a fast deployment of HTML files for annotation. Each XHTML page abides by specific microformat vocabularies to markup the content. XSLT[2] can be used to convert XHTML structure into RDF/XML. GRDDL[3], a W3 standard enables processors to locate the XSLT transformation for microformats.

If we want to add semantic information to simple data structures like people or calendar, then microformats can be used but microformats are difficult to scale. Microformats does not interfere with one another, slowing the adoption of new vocabularies. If a single HTML content contains different microformat data for different vocabularies, it becomes difficult to embed. Microformats are not aW3C standard. They are defined and maintained by an open community [4].

B. RDFa:

More complex approach for semantic markup can be provided by RDFa. RDFa is defined by the world wide consortium. RDFa provides its own set of attributes that are used in conjunction with XHTML constructs. Hence RDFa provides flexibility to add any information on to the web page. This is the reason for the name RDFa which means RDF in HTML attributes. Unlike microformat+GRDDL approach, RDFa specification gives a mapping from XHTML+RDFa file to RDF. RDFa+XHTML can be seen as RDF serialization format, beside RDF/XML.

Like microformats, RDFa use XHTML structure as a framework to include RDF information. This allows RDF to be carried along with HTML. By embedding semantic information into web pages, there is no need for screen scraping.

C. Microdata:

Microdata uses a number of attributes found in microformats. It is proposed and maintained by W3C. Microdata annotate web pages using DOM attributes not XML. Microdata vocabulary can be defined by anyone and can be used for embedding custom properties in their web pages.

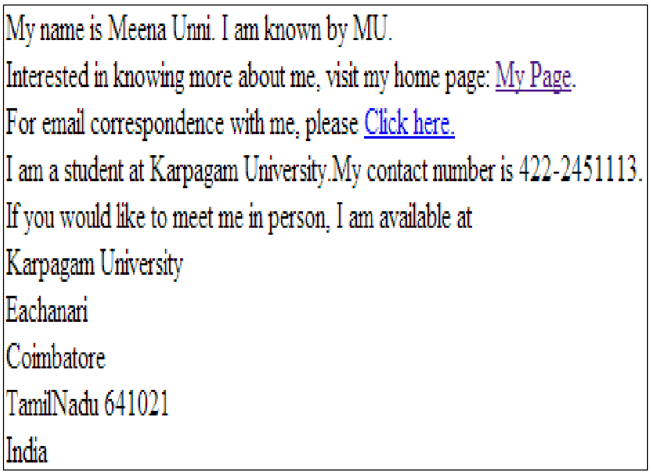
III. APPLICATION OF TECHNOLOGIES

Below sections presents usage of microformat and RDFa and microdata in XHTML documents. These technologies can be used for other XML[5] dialects too.

A. Microformats:

Microformats embed semantic data into HTML content by reusing HTML attributes. Each application is associated with its own vocabulary and syntax. To add some semantic data about people, hcard microformat which gives a group of constructs to mark up the content can be used. It consists of a root called vcard and a collection of properties, such as name, address, tel etc. Microformat vocabularies are designed by translating existing vocabularies to the microformats approach. Strings values are used to determine semantic information in attribute values. Designing new microformats involve defining these strings and giving appropriate meaning to it.

Consider an example code in HTML with its output page.



My name is Meena Unni. I am known by MU.

Interested in knowing more about me, visit my home page: My Page.

For email correspondence with me, please Click here.

I am a student at Karpagam University.My contact number is 422-2451113.

If you would like to meet me in person, I am available at

Karpagam University

Eachanari

Coimbatore

TamilNadu 641021

India

Figure 2: HTML Output Page

```
<html>
<title> Testing Page </title>
<div>
 My name is Meena Unni. I am known by MU. <br>
Interested in knowing more about me, visit my home page:
 <a href="http://mu123.site90.net">My Page</a>.<br>
 For email correspondence with me, please <a href=
"mailto:meena@ku.com">Click here.</a><br>
 I am a student at Karpagam University.My contact number is
 422-2451113.<br>If you would like to meet me in person,
 I am available at<br>
 Karpagam University<br>
 Eachanari<br>
 Coimbatore <br>
 TamilNadu 641021<br>
 India<br>
</div>
</html>
```

Figure 3: HTML Code

Below gives the same XHTML page enhanced using microformats. Output of the page is similar to that shown in figure 1.

```
<title> Testing Page </title>
<div id="hcard-Meena" class="vcard">
<img class="photo" src="india-flag.gif" height="25" width="45" />

<a class="n fn url" href="http://mu123.site90.net"><br>
My name is
<span class="given-name">Meena</span>
<span class="family-name">Unni</span>.</a> I am known by
<span class="nickname">MU.</span><br> Interested in knowing more
about me,
 visit my home page:<a class="url" href=""http://mu123.site90.net">My
Page</a>
<br>For email correspondence with me, please
<a class="email" href="mailto:meena@ku.com">
Click here.</a><br>
I am a <span class="role"> student </span> at Karpagam University.
<span class="tel">
<span class="type"> My contact number is </span>
<span class="value">422-2451113</span>
<br>
If you would like to meet me in person, I am available at <br>
<div class="fn org">Karpagam University</div>
<div class="adr">
<div class="street-address">Eachanary</div>
<div class="locality">Coimbatore</div>
<span class="region">TamilNadu</span>
<span class="postal-code">641021</span>
<div class="country-name">India</div>
</div>
```

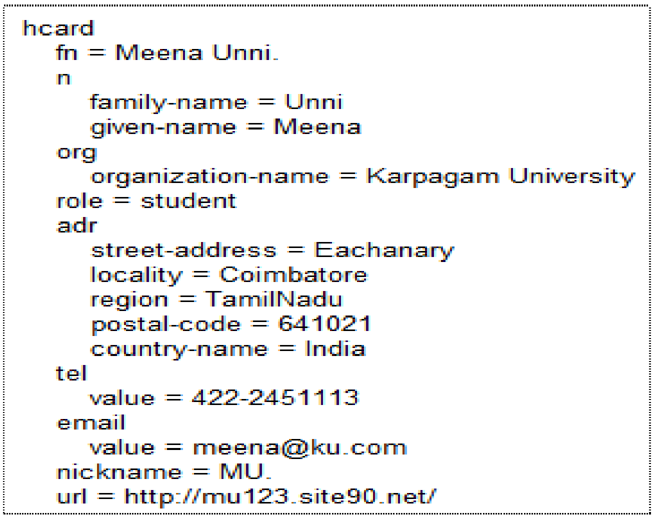
Figure 4: XHTML embedded with microformats

Above shows the previous web page enhanced using microformat. Information is added to the XHTML page but is not visible on the browser screen.

The issue with Microformat tag is that they are independent of another. Developing an application to integrate different tags means understanding the different tags and how they are coded and parsed. Mocroformat community makes sure that there are no conflicts between tags. This requires centralized control, which would prevent decentralized innovation.

Microformats are simple to use. Main issue with microformat is that it supports less number of attributes. Microformats can be used to markup people, events, organization, reviews but there is no attributes to markup other domain like movie or songs. Microformats are obscure and hard to read.

The extracted information from the above XHTML code enriched with microformat is given below.



```
hcard
   fn = Meena Unni.
   n
      family-name = Unni
      given-name = Meena
   org
      organization-name = Karpagam University
   role = student
   adr
      street-address = Eachanary
      locality = Coimbatore
      region = TamilNadu
      postal-code = 641021
      country-name = India
   tel
      value = 422-2451113
   email
      value = meena@ku.com
   nickname = MU.
   url = http://mu123.site90.net/
```

Figure 5: Extracted data from the page

B. RDFa:

RDFa like microformats use XHTML attributes to express additional content, and depend on external processing to retrieve necessary information. RDFa publishers can reuse the defined vocabulary whereas microformats must redefine the vocabulary.

One of the widely used vocabulary in RDF applications is based on data-vocabulary, Friend-of-a-Friend. Vocabulary can be used to describe personal data and relationships between people. If vocabulary is very large, using microformat would be complex for semantic annotation.

RDFa can handle multiple vocabularies well as it is based on the usage of Uniform Resource Indicators (URIs). RDFa has a namespace like mechanism to abbreviate URIa to strings like foaf:name, foaf:address. These prefixes are stated using XMLns. These compact URIs are called CURIEs.

Below shows HTML file after adding RDFa attribute.

```
<title> Testing Page </title>
<img src="india-flag.gif" alt="Rainfall 1900-1999" height=40 width=40/>.
<div xmlns:v="http://xmlns.com/foaf/0.1/" typeof="v:Person">
My name is <span property="v:name">Meena Unni</span>,
I am known by <span property="v:nickname">MU</span>.<br>
Interested in knowing more about me, visit my home page :
<a href="http://mu123.site90.net" rel="v:url">My Page</a>.<br>
For email correspondence, please
<a class="email" href="mailto:meena@ku.com">Click here </a>
<br>  I am a  <span property="v:title">student</span>
at <span property="v:affiliation">Karpagam University</span>.
My contact number is <span property="v:tel">422-2451113 </span>
</div>
If you would like to meet me in person, I am available at
<br>Karpagam University <br>
  <span rel="v:address">
    <span typeof="v:Address">
      <div property="v:street-address">Eachanary</div>
       <div property="v:locality">Coimbatore</div>
       <span property="v:region">TamilNadu</span>
 <span property="v:postal-code">641021 </span>
       <div property="v:country-name">India </div>
  </span>
  </span>
```

Figure 6: Code showing RDFa embedded in XHTML

RDFa is a thin layer of markup that you can add to your web pages to make them understandable for machines as well as people. By adding it, browsers, search engines, and other software can understand more about the pages, and in so doing offer more services or better results for the user [6] Unlike microformats which reuse the existing class attribute on most HTML tags, RDF provides a set of new attributes that can be used to carry the added markup data [7]. Microformats reuse existing class attribute on HTML tags whereas RDFa provides a set of new attributes to markup the data. Extracted data from the page is similar to that shown for microformats.

C. Microdata:

Microdata enriches microformats attributes. Microdata uses itemprop attributes instead of class. Unlike RDFa, microdata does not support datatypes of literals, and XML literals.

Below shows the same XHTML file embedded with microdata.

```
<div itemscope itemtype="http://data-vocabulary.org/Person">
<img src="india-flag.gif" itemprop="image" height="25" width="40"
/><br>

  My name is <span itemprop="name">Meena Unni</span>
  I am known by <span itemprop="nickname">MU</span>. <BR>
 Interested in knowing more about me, visit my home page:
  <a href="http://mu123.site90.net" itemprop="url">My Page</a>

<br>
For email correspondence with me, please <a
href="mailto:meena@ku.com">Click here.</a><br>
I am a <span itemprop="title">student</span> at <span
itemprop="affiliation"> Karpagam University </span>.My contact number
is
<span itemprop="telephone">422-2451113</span>
<br>
If you would like to meet me in person, I am available at<br>
Karpagam University<br>
<span itemprop="address" itemscope
     itemtype="http://data-vocabulary.org/Address">
<div itemprop="street-address"> Eachanari</div>
<div itemprop="locality"> Coimbatore</div>
<span itemprop="region"> TamilNadu</span>
<span itemprop="postal-code">641021 </span>
<div itemprop="country-name"> India</div>
</span>
```

Figure 7: Code showing microdata embedded into XHTML

Extracted data from the page enhanced using microdata is similar to that shown for microformats.

IV. CONCLUSION AND FUTURE WORK

Vision of semantic web is that the machines should be able to read the web page independently and its success depends on deployment. Microformats, RDFa and microdata is a major step towards this deployment. Improved search is the primary advantage of semantic markup, which pulls information from the web pages and presents it in a way useful to users. An application which understands these technologies can perform much more complex tasks than that can done by screen scraping.

Another advantage of these technologies is that they can be included with XHTML. These technologies makes web users connect data, search and extract information. This paper presents a novel approach to embedding data using semantic markup. The paper also gives example to illustrate that using semantic annotation, does not change the look and feel of the document. It adds only meaning to data in such a way that the machine can read and interpret the information present in the web page.

V. REFERENCES

[1]. T. Berners-Lee, J. Hendler, and O. Lassila. "The Semantic Web", "Scientific American", 284(5):34–43, May 2001.

[2]. J. Clark, (ed.): " XSL Transformations (XSLT)", W3C recommendation. http://www.w3.org/TR/ xslt (1999)

[3]. D. Connolly, (ed.): "Gleaning Resource Descriptions from Dialects of Languages (GRDDL)", W3C recommendation. http://www.w3.org/TR/grddl/ (2007)

[4]. Microformats. http://microformats.org

[5]. Michael C., DacontaLeo J. Obrst, Kevin T. Smith, "The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management", 2003

[6]. A. Rajendra, "Semantic Markup Report , Microformats, RDFa, GRDDL, Microdata and OGP", "NCE Tourism Fjord Norway"

[7]. J. Simpson, "Microformats vs. RDF: How Microformats relate to the Semantic Web", April, 2009, http://www.semanticfocus.com/blog/entry/title/microformats-vs-rdf-how-microformats-relate-to-the-semanticweb