

Website Parse Template v1.0

Introduction

Existing web crawling technologies assume content extraction directly from web pages without basing on keywords declared in HTML codes. The key reason is that web publishers usually define keywords different from the actual content. The same situation is with deployment of RDF, because there is no guarantee for web crawlers that the information included in RDF file fully corresponds to actual content. Moreover, RDF description prepared for specific web page provides information about the content and does not include any information about content allocation on that web page.

To escape the mismatch problems described above, web crawlers are forced to check RDF compliance with structured content for each targeted web page which is associated with the lack information on web page structured content. Website Parse Template facilitates solution of this problem by providing web page HTML structure description for a single or group of similarly structured pages. Website Parse Template (WPT) allows web publishers to define references to specific HTML elements together with web page content description represented in any supported format including RDF.

Overview

Website Parse Template (WPT) is an XML based open format which provides web crawlers with additional proper information on web page HTML structure and content. WPT is compatible with XML schemas, such as RDF and OWL.

Website Parse Template uses XML tags of Open Web vocabulary which is being declared as an XML namespace: `xmlns:ow="http://www.omfica.org/schemas/ow/0.9"`.

Website Parse Template begins with opening `<ow:wpt>` tag and ends with closing `</ow:wpt>` tag. Single Website Parse Template is referred to the same host, while single host may have several Website Parse Templates describing its HTML structure. It is required to specify host name the Website Parse Template is for and declare the namespace within `<ow:wpt>` tag (see example below).

Example 1. Website Parse Template frame

```
<?xml version="1.0" encoding="UTF-8"?>
<ow:wpt xmlns:ow="http://www.omfica.org/schemas/ow/0.9"
        ow:host="http://www.example.com">
.....
</ow:wpt>
```

Website Parse Template consists of following sections:

- **Templates** is a mandatory section, which contains web pages' HTML structure and content description.
- **URLs** is an optional section, which links URL Patterns for groups of web pages to specified Templates.
- **Ontology** is an optional section which defines concepts and relations used in website.

Website Parse Template v1.0

Templates

Templates section describes web page's HTML structure by making references to corresponding HTML elements of specific web page.

Template starts with opening `<ow:template>` tag and ends with closing `</ow:template>` tag. It is required to specify unique template name within `<ow:template>` tag and define URL which complies with specific template.

Template consists of blocks corresponding to each structural element of specific web page. Each template must contain at least one block. Block makes reference to appropriate HTML element through one or any combination of following reference methods: TagID, XPath and Pattern. Each block must start with opening `<ow:block>` tag and correspondingly ends with closing `</ow:block>` tag. It is required to indicate specific HTML element reference(s) within block's opening tag.

Example 2. Block reference methods

```
<ow:template ow:name="Template Example" ow:url="http://www.example.com/index.php">
.....
<ow:block ow:tagid="ex1" ow:xpath="/html/body/div/div" ow:pattern="content ([a-z]*)">
..... //content description
</ow:block>
<ow:block ow:tagid="ex2">
..... //content description
</ow:block>
<ow:block ow:xpath="/html/body/div/div/table/tr[1]/td">
..... //content description
</ow:block>
.....
</ow:template>
```

Each block contains specific HTML element's content description represented solely or within another block. Embedded blocks are used to describe specific HTML element ("parent block") which includes one or more elements ("child block").

Example 3. Embedded blocks

```
<ow:template ow:name="Template Example" ow:url="http://www.example.com/index.php">
.....
<ow:block ow:xpath="/html/body/div/div">
  <ow:block ow:xpath="/html/body/div/div/table/tbody/tr[1]/td">
    ..... //content description
  </ow:block>
</ow:block>
.....
</ow:template>
```

Website Parse Template v1.0

Content description can be provided using concepts defined in Ontology section or any supported format/language: RDF, CWL, etc. It is required to declare namespaces of used XML schema(s) within <ow:wpt> tag and ontology name within <ow:template> tag.

Example 4. Content description instances

```
<ow:wpt xmlns:ow="http://www.omfica.org/schemas/ow/0.9"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  ow:host="http://www.example.com">
  .....
  <ow:template ow:name="Template Example" ow:url="http://www.example.com/index.php"
  ow:ontology="ontology_example">
  .....
  //Content description using defined concepts
  <ow:block ow:tagid="ex1" ow:xpath="/html/body/div/div" ow:pattern="Wellcome (user.name[[A-Za-z]*])">
    ontology_concept
  </ow:block>
  //Content description using RDF syntax
  <ow:block ow:tagid="ex2">
    <rdf:Description rdf:about="http://www.example.com/index.php">
      .....
    </rdf:Description>
  </ow:block>
  //Content description using CWL.unl
  <ow:block ow:xpath="/html/body/div/div/table/tr[1]/td">
    {cwl.unl}
    .....
    {/cwl.unl}
  </ow:block>
  .....
</ow:template>
.....
</ow:wpt>
```

If the web page contains listed or structured repeatable content it can be represented as a single entry by specifying block type as repeatable. For example, if specific HTML element repeats several times as a list it can be described as a single repeatable block instead of specifying blocks for each listed element. In most cases repeatable blocks are child blocks embedded within parent block – another HTML element.

Example 5. Repeatable content representation

```
<ow:template ow:name="Template Example" ow:url="http://www.example.com/index.php">
  .....
  <ow:block ow:xpath="/html/body/div/table/tbody/tr/td[2]" ow:type="repeatable">
    ..... //content description
  </ow:block>
  <ow:block ow:xpath="/html/body/div/div">
    <ow:block ow:xpath="/html/body/div/div/table/tbody/tr[1]/td" ow:type="repeatable">
      ..... //content description
    </ow:block>
  </ow:block>
  .....
</ow:template>
```

Website Parse Template v1.0

If the web page contains listed or structured repeatable content it can be represented as a single entry by specifying block type as repeatable. For example, if specific HTML element repeats several times as a list it can be described as a single repeatable block instead of specifying blocks for each listed element. In most cases repeatable blocks are child blocks embedded within parent block – another HTML element.

Example 6. Pattern references

```
<ow:template ow:name="Template Example" ow:url="http://www.example.com/index.php" ow:ontology="ontology_example">
.....
<ow:block ow:pattern="<td bgcolor=\\"FFFFFF\\" class=\\"small\\"><a href=\\"/ar-(artist.id[[0-9]*])---
                    (artist.name[[A-Za-z]*)\\" class=\\"small\\"><b>(artist.name[[A-Za-z]*])" ow:type="repeatable">
..... //content description
</ow:block>
<ow:block ow:pattern="The Best Music Blogs on the Web">
..... //content description
</ow:block>
<ow:block ow:pattern="Wellcome (member.name[[A-Za-z0-9]*])! Your member ID is (member.id[[0-9]*]). >
..... //content description
</ow:block>
.....
</ow:template>
```

Single template may describe single web page or a group of similarly structured web pages. In the examples above template refers to a single web page (e.g. ow:url="http://www.example.com/index.php"). In case of describing group of similarly structured web pages via single template it is necessary to define URL pattern covering all of those pages (e.g. ow:url="http://www.example.com/(page.name[[a-z]*]).php").

Example 7. Template for a Single Artist Page on Yahoo! Music

```
<?xml version="1.0" encoding="UTF-8"?>
<ow:wpt xmlns:ow="http://www.omfica.org/schemas/ow/0.9"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        ow:host="http://music.yahoo.com">
.....
<ow:template ow:name="Artist Page on Yahoo! Music"
            ow:url="http://music.yahoo.com/ar-(artist.id[[0-9]*])---(artist.name[[A-Z,a-z,-,0-9]*])"
            ow:ontology="general">
<ow:block ow:tagid=="yent-uhdr">Menu</ow:block>
<ow:block ow:xpath="/html/body/div[2]/div/div/div[3]/div/a/span">Logo</ow:block>
<ow:block ow:xpath="/html/body/div/div">Advertisement</ow:block>
<ow:block ow:xpath="/html/body/div[3]/table/tbody/tr/td[2]/div/h1">artist.name</ow:block>
<ow:block ow:tagid="art_img">artist.image</ow:block>
<ow:block ow:tagid="biography">artist.bio</ow:block>
<ow:block ow:xpath="/html/body/div[3]/table/tbody/tr/td[2]/table/tbody/tr[22]">artist.album</ow:block>
<ow:block ow:xpath="/html/body/div[3]/table/tbody/tr/td[2]/table/tbody/tr[10]">artist.track</ow:block>
<ow:block ow:xpath="/html/body/div[3]/table/tbody/tr/td[2]/table/tbody/tr[13]">artist.video</ow:block>
</ow:template>
.....
</ow:wpt>
```

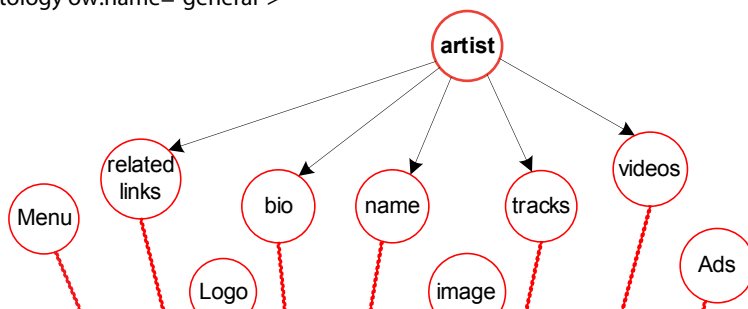
Website Parse Template v1.0

See visual representation of Website Parse Template in the figure 1 below.

Figure 1. WPT visual representation

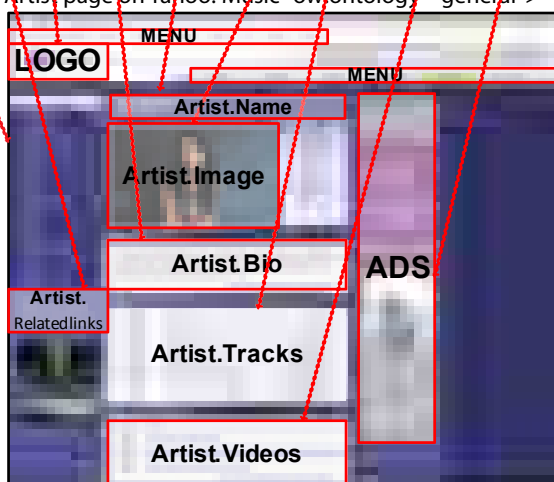
```
<ow:wpt ow:host="music.yahoo.com">
```

```
<ow:ontology ow:name="general">
```



```
</ow:ontology>
```

```
<ow:template ow:name="Artist page on Yahoo! Music" ow:ontology="general">
```



```
</ow:template>
```

```
<ow:urls ow:name="Artist page on Yahoo! Music" ow:template="Artist page on Yahoo! Music">
```

```
<ow:url>http://music.yahoo.com/ar-(artist.id[[0-9]*])--(artist.fullname[[A-Za-z0-9-]*])</ow:url>
```

```
<ow:url>http://music.yahoo.com/ar-8206256---Amy-Winehouse</ow:url>
```

```
</ow:urls>
```

```
</ow:wpt>
```

URLs section

This section defines the URLs/URL patterns of web pages described in Templates section. This section is mandatory if the templates do not define URLs/URL patterns of web pages.

In accordance with Templates section this section also may consist of several blocks/units. Either of those blocks starts with <ow:urls> tag and ends with </ow:urls> tag.

Website Parse Template v1.0

Example 8. URL patterns

```
<ow:urls ow:name="Artist page on Yahoo! Music" ow:template="Artist page on Yahoo! Music">  
  <ow:url>http://music.yahoo.com/ar-8206256---Amy-Winehouse</ow:url>  
  <ow:url>http://music.yahoo.com/ar-([artist.id[0-9]*])---(artist.name[[A-Za-z0-9-]*])</ow:url>  
</urls>
```

As a URL block's name can be chosen any string, but for the template it is necessary to indicate specific template name described in previous section.

RegExp specifications are used for URL patterns descriptions. The URL pattern provided in Example 8 also includes the represented real URL. The concepts necessary for URL pattern definition (such as "id" and "name") are to be defined in Ontology section.

WPT Ontology

Ontology section contains enumeration and definition of all concepts used in website. Listed concepts must be enclosed between `<ow:ontology>` and `</ow:ontology>` tags. It is required to specify the ontology name (any rational string) within `<ow:ontology>` tag. WPT allows using of either OWL or WPT Ontology language for concepts definition. The main difference between those languages is that WPT Ontology language provides simplified syntax for concepts and relations definition.

Example 9. "artist" concept definition using WPT Ontology language

```
<ow:ontology ow:name="general">  
  <ow:concept ow:name="artist">  
    <ow:inherit>person</ow:inherit>  
    <ow:has>name</ow:has>  
    <ow:has>album</ow:has>  
    <ow:has>track</ow:has>  
    <ow:has>image</ow:has>  
    <ow:has>bio</ow:has>  
    <ow:has>video</ow:has>  
    <ow:has>id</ow:has>  
  </ow:concept>  
  <ow:concept>logo</ow:concept>  
  <ow:concept>menu</ow:concept>  
  <ow:concept>advertisement</ow:concept>  
</ow:ontology>
```

Each concept definition starts with `<ow:concept>` tag and ends with `</ow:concept>` tag. `<ow:inherit>` tag shows inheritance relations and `<ow:has>` tag shows attributable relations between two concepts. Either of defined concepts has default attribute - object identifier (**id**) to be used by web crawlers to co-ordinate the same object's attributes used in different pages of the same website.