External data in Semantic MediaWiki

Yaron Koren

Semantic wiki mini-series session 5

February 12, 2009

The problem

How can a Semantic MediaWiki wiki use data from outside sources?

Three kinds of sources

1) Sites that have their own web-based API for accessing data

2) Other Semantic MediaWiki sites

3) Everything else – standalone databases, spreadsheets, etc.

A solution: "External Data" extension

Written by Yaron Koren and Michael Dale

http://www.mediawiki.org/wiki/Extension:External_Data

External Data - defines three MediaWiki "parser functions"

#get_external_data – retrieves the contents of a URL that holds data in either CSV, JSON or XML formats, and sets it to local variables

#external_value – displays the value of a single local variable

#for_external_data – loops through the rows of a retrieved table of variables

Data type #1: Site with an API

The URL "http://fruits.com/api?fruit_name=Apple" (from a non-wiki site) contains the following XML data:

```
<fruit><name>Apple</name><color>Red</color>
<size>Medium</size></fruit>
```

The wiki page then contains this call:

```
{{#get_external_data:http://fruits.com/api?
fruit_name=Apple|xml|fruit_color=color|fruit_size=size}}
```

...this call sets the local variables "fruit_color" and "fruit_size" to be "Red" and "Medium", respectively.

Data type #1: Site with an API (cont.)

Then the page contains the following:

```
An apple has color {{#external_value:fruit_color}}, and
size {{#external_value:fruit_size}}.
```

The page could even contain:

```
An apple has color [[Has color::
{{#external_value:fruit_color}}]], and size [[Has size::
{{#external_value:fruit_size}}]].
```

...and thus external data could be stored semantically, so it can be queried like any local data.

Data type #1: Site with an API (cont.)

Or, we can get data from the URL “http://fruits.com/api?get_all_fruits”, which contains data on all fruits. The wiki page would then contain this call:

```
{{#get_external_data:http://fruits.com/api?get_all_fruits|
xml|fruit_name=name|fruit_color=color|fruit_size=size}}
```

Then it could call the following:

```
{{#for_external_table:The fruit {{{fruit_name}}} has color
{{{fruit_color}}}, and size {{{fruit_size}}}.}}
```

This would print that statement for every retrieved row.

(We could similarly use #for_external_table to print out a table.)

Data type #2: SMW site

SMW's inline queries support "CSV", and (as of very recently) “JSON” formats. The URL

http://semanticweb.org/wiki/Special:Ask/-5B-5BGermany-5D-5D/-3FHas_capital/-3FPopulation/format%3Dcsv/sep%3D,

contains the following text:

```
Germany,Berlin,82411000
```

On a wiki page, we can call:

```
{{#get_external_data:http://semanticweb.org/wiki/Special:As
k/-5B-5BGermany-5D-5D/-3FHas_capital/-3FPopulation/format
%3Dcsv/sep%3D,|csv|capital=1|population=2}}
```

Data type #3: Inaccessible data

- The hardest type of data to deal with, but also by far the most common.

- So far, the solution has usually been to enter data into the wiki as a large set of wiki pages, using either template calls or semantic property calls. Unfortunately, maintenance is difficult.

- If only there were a way to easily create an API for accessing that data...

The External Data solution

Create a wiki page for each table of data, in CSV format.

Example, for a page called “World leaders data”:

```
Name,Country,Start year,End year
Abdallah al-Adil,Morocco,1224,1227
Abdullah Gul,Turkey,2002,2003
Adalbert,Italy,950,963
...
```

The External Data solution (cont.)

'Special:GetData' can then serve as a mini-API for getting values from that page.

Example URL, to get all the leaders of France:

http://mywiki.com/Special:GetData/World_leaders_data?Country=France

The External Data solution (cont.)

- The data is stored on the wiki, but in an easily-manageable form; no need for maintaining many pages with many template or property calls.

- This approach more closely resembles that of a “document-oriented database” than SMW does

“Data type #4”: User-generated data

Does it ever make sense to use the External Data approach even for user-generated data, internal to the wiki?

One obvious usage: table data on a page – SMW does not allow this yet, due to lack of support for n-ary relations/internal objects

Exit question

Will there be uses for the data-page approach for user-generated data even when SMW has complete support for tables?

A possible example: a multi-language wiki – data is stored in one page holding CSV and accessed similarly by all language versions of the wiki.

Exit question #2

Hey, isn't Wikipedia a multi-language wiki?