Hooking up Semantic MediaWiki with external tools via SPARQL

The Use case behind the RDFIO extension

RDFIO Extension: Google Summer of Code project in 2010
   Mentor: Denny Vrandecic <http://simia.net/>

SMW + Bioclipse integration: MSc Thesis project in Bioclipse group, Inst. of Pharmaceutical Biosciences, Uppsala University, Sweden.
   Supervisor: Egon Willighagen <http://chem-bla-ics.blogspot.com>
   Samuel Lampa <http://saml.rilspace.org>, SMWCon 23 Sep 2011
This talk

● The motivation and use case behind RDFIO
● Looking for feedback on design decisions
● Outlook and Plans
● Possible collaborations / interactions?
The Problem

• Need to combine:
  ○ Manual schema exploration, as knowledge grows
  ○ Automated data generation, for integrating data from literature or own experimental or computational results
  ○ Community collaboration

• Thus need something that supports all in same framework
Solution: Bioclipse + SMW

Bioclipse as automated or interactive batch writer
Bioclipse Bioinformatics Workbench
Bioclipse Shortfacts

- Open Source
- Based on Eclipse - Modular, Solid
- Cheminformatics/Bioinformatics Tools
  - Jmol, JChemPaint, etc
- Scripting (JavaScript)
- Semantic Web Support
- Core team at Uppsala University
- [http://bioclipse.net](http://bioclipse.net)
SMW Limitations

- No (prior) RDF import
- No obvious way to choose wiki titles for URIs (Problem 1)
- RDF Export format not same as import format ... is hard-coded, using URIs such as:
  
  \[
  \text{http://example.com/wiki/Special:URIResolver-3ASomeEntity}
  \]

  (Problem 2)
So, needed a new extension: RDFIO
RDFIO Extension Short facts

● Google Summer of Code project 2010
  ○ Mentor: Denny Vrandecic

● Dependencies:
  ○ Semantic MediaWiki
  ○ SMWWriter (Denny Vrandecic)
  ○ POM (Sergey Chernychev)
  ○ ARC2 RDF Lib (Benjamin Nowack)

● Current version: 0.5
  ○ Needs bugfixing
  ○ Needs making SMW 1.6 compatible

● http://www.mediawiki.org/wiki/Extension:RDFIO
RDFIO Data Flow

SPARQL+ interface -> RDF Import form

RDFIO
SMW Writer
POM Extension

Triple store -> MediaWiki DB
Problem 1: Choosing wiki page titles for RDF entities
Example of Experimental Data

Molecule

Spectrum

Peak

has

Peak

has

Peak

has

Peak

has

Peak

has

Peak

has

Peak

has

Peak

has

Peak

has
<http://[...]/nmrshiftdb/?moleculeId=234>
dc:title "warburganal";
chem:casnumber "62994-47-2";
nmr:moleculeId "234";
nmr:hasSpectrum <http://[...]/nmrshiftdb/?spectrumId=4735>;
<http://[...]/nmrshiftdb/?spectrumId=4735> nmr:field "50";
nmr:hasPeak <http://[...]/nmrshiftdb/?s4735p0>,
<http://[...]/nmrshiftdb/?s4735p1>,
<http://[...]/nmrshiftdb/?s4735p2>,
<http://[...]/nmrshiftdb/?s4735p3>;
nmr:solvent "Chloroform-D1 (CDCl3)";
nmr:spectrumId "4735";
nmr:spectrumType "13C";
nmr:temperature "298".
<http://[...]/nmrshiftdb/?s4735p1>
nmr:hasShift 18.3;
a nmr:peak.
<http://[...]/nmrshiftdb/?s4735p2>
nmr:hasShift 22.6;
a nmr:peak.
<http://[...]/nmrshiftdb/?s4735p3>
nmr:hasShift 26.5;
a nmr:peak.
Problem: URIs not nice as titles
So, how to choose the wiki page title then?
Let's look what we have in the RDF, that might be used ...
Wiki page title for the Molecule?

<http://[...]/nmrshiftdb/?moleculeId=234>
dc:title "warburganal";
chem:casnumber "62994-47-2";
nmr:moleculeId "234"
  nmr:hasSpectrum <http://[...]spectrumId=4735>;
Wiki page title for the Molecule?

<http://[...]/nmrshiftdb/?moleculeId=234>
dc:title "warburganal";
chem:casnumber "62994-47-2";
nmr:moleculeId "234";
nmr:hasSpectrum <http://[...]/spectrumId=4735>;}
Wiki page title for the Spectrum?

<http://[...]nmrshiftdb/?spectrumId=4735>
nmr:field "50";
nmr:hasPeak <http://[...]nmrshiftdb/?s4735p0>,
   <http://[...]nmrshiftdb/?s4735p1>,
   <http://[...]nmrshiftdb/?s4735p2>,
   <http://[...]nmrshiftdb/?s4735p3>;
nmr:solvent "Chloroform-D1 (CDCl3)";
nmr:spectrumId "4735";
nmr:spectrumType "13C";
nmr:temperature "298".
Wiki page title for the Spectrum?

<http://[...]nmrshiftdb/?spectrumId=4735>
nmr:field "50";
nmr:hasPeak <http://[...]nmrshiftdb/?s4735p0>,
<http://[...]nmrshiftdb/?s4735p1>,
<http://[...]nmrshiftdb/?s4735p2>,
<http://[...]nmrshiftdb/?s4735p3>;
nmr:solvent "Chloroform-D1 (CDCl3)"

nmr:spectrumId "4735";
nmr:spectrumType "13C";
nmr:temperature "298".
Wiki page title for the Spectrum?

<http://pele.farmbio.uu.se/nmrshiftdb/?s4735p1>
  nmr:hasShift 18.3;
  a nmr:peak.
Wiki page title for the Spectrum?

<http://pele.farmbio.uu.se/nmrshiftdb/?s4735p1>  
nmr:hasShift 18.3;  
a nmr:peak.
So ...

- So, we might find useful titles in:
  - Dublin Core fields (small vocabulary for cataloging)
  - Custom identifiers
  - (parts of) URIs

- So, we need a flexible configuration system such as:
  - A system-wide defaults in `LocalSettings.php`, AND
  - Interactive configuration in RDF Import form

(See following slides ...)
In LocalSettings.php ...

/**
 * Used to generate "Pseudo-NameSpaces"
 */
$rdfiogBaseURIs = array(
    "http://example.org/someOntology#" => "ont1",
    "http://example.org/anotherOntology#" => "ont2"
);

/**
 * "Properties" to be used as Wiki Title
 */
$rdfiogPropertiesToUseAsWikiTitle = array(
    'http://semantic-mediawiki.org/swivt/1.0#page',
    'http://www.w3.org/2000/01/rdf-schema#label',
    'http://purl.org/dc/elements/1.1/title',
    'http://www.w3.org/2004/02/skos/core#preferredLabel',
    'http://xmlns.com/foaf/0.1/name',
    'http://www.nmrshiftdb.org/onto#spectrumId'
);
... and in RDF Import Screen (1/2)
... and in RDF Import Screen (2/2)
Result: *Problem 1 solved!*

Page get sensible titles

*(See screenshot on next slide)*
**Spectrum:4735**

### Facts about Spectrum:4735

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nmr:field</td>
<td>50</td>
</tr>
<tr>
<td>Nmr:solvent</td>
<td>Chloroform-D1 (CDCl3)</td>
</tr>
<tr>
<td>Nmr:spectrumId</td>
<td>4735</td>
</tr>
<tr>
<td>Nmr:spectrumType</td>
<td>13C</td>
</tr>
<tr>
<td>Nmr:temperature</td>
<td>298</td>
</tr>
</tbody>
</table>
Problem 2: Export RDF in the format it was imported
(... not the SMW hard-coded format)
Solution: Storing Original URIs
Result: *Problem 2 solved!*
Can Query/Export in original RDF format
SPARQL Endpoint

Enter SPARQL query:

```
@PREFIX w: <http://drugmet.rilspace.org/wiki/Special:URIResolver/> .

SELECT ?p ?o
LIMIT 25
```

- Query by original URIs: ✔️
- Output original URIs: ✔️

Output format: HTML

(RDF/XML requires creating triples using CONSTRUCT)
So now we have solved the problems:

- Choosing Wiki Titles
- Export in same format as import

... so now we can go ahead and hook up SMW with an external tool ....
Talking to SMW from Bioclipse scripts

```javascript
var wikiURL = "http://drugmet.rilspace.org/wiki/";

/* Editing SMW facts */
smw.addTriple("w:Caffeine", "w:is_a", "w:Molecule", wikiURL);
smw.removeTriple("w:Caffeine", "w:is_a", "w:Molecule", wikiURL);

/* Downloading RDF for local querying */
rdfStore = rdf.createInMemoryStore();
rdfStore = smw.getRDF( wikiURL );
result = rdf.sparql( rdfStore,
  "SELECT DISTINCT ?pred WHERE { ?subj ?pred ?obj } LIMIT 10"
)

/* Show some output */
js.print( result );
```
Talking to SMW from Bioclipse (contd.)

...
So ... the use case is now possible ...
Outlook

- Proof of concept is there

- Things to fix before useful:
  - SMW 1.6 support
  - Some bugfixes

- Further wishlist:
  - Support more triple stores
  - Skip ARC2 PHP RDF library dependency?
  - SPARQL UPDATE instead of ARC2's SPARQL+
  - "Output as Original URIs" also in RDF Batch export
  - Improve performance
Thank you!

RDFIO Extension: Google Summer of Code project in 2010
Mentor: Denny Vrandecic <http://simia.net/>

Bioclipse + SMW integration: MSc Thesis project in Bioclipse group, Inst. of Pharmaceutical Biosciences, Uppsala University, Sweden.
Supervisor: Egon Willighagen <http://chem-bla-ics.blogspot.com>
Samuel Lampa <http://saml.rilspace.org>, SMWCon 23 Sep 2011