Towards Querying Bioinformatic Linked Data in Natural Language

Anca Marginean si Oana Marc

September 6, 2013
1. Context
   - Linked Data - the concept
   - Existing approaches for querying

2. System overview
   - System’s modules
   - Bio2RDF. DrugBank
   - Pattern 1 - properties of a certain drug
   - Pattern 2 - properties of classes of drugs
   - Other patterns

3. Conclusions
“If HTML and the Web make all online documents look like one huge book, RDF, schema, and inference languages will make all the data in the world look like a huge database.” Tim Berners-Lee
Linked Data

Triplestores - sentences as triplets

Subject  Predicate  Object
or  Entity  Property  Value

SPARQL language

What are the birth date and the directed movies of Roman Polanski?

PREFIX dbpedia2: <http://dbpedia.org/property/>
SELECT ?date ?movie
WHERE {
}
Linked Data

**Triplestores - sentences as triplets**

Subject  Predicate  Object

or  Entity  Property  Value

**SPARQL language**

What are the birth date and the directed movies of Roman Polanski?

```sparql
PREFIX dbpedia2: <http://dbpedia.org/property/>
SELECT ?date ?movie
WHERE {
}

PREFIX dbpedia: <http://dbpedia.org/>
```
Linked Data - existing approaches for querying

Querying RDF triplestores
- desktop clients: Explorator, RDF-gravity, Twinkle
- SPARQL endpoints
- assisted incremental query building
- CNL - Controlled Natural Language:
  - Montague grammar: each rule for syntactic structure is decorated by a \( \lambda \)-term

Translation of Natural Language to meaning representation language:
- CHILL - Prolog based language
- WASP - functional language for Geoquery
System overview

- Bio2RDF, DrugBank
- Pattern 1: properties of a certain drug
- Pattern 2: properties of classes of drugs
- Other patterns

Anca Marginean și Oana Marc
Towards Querying Bioinformatic Linked Data in Natural Language
Bio2RDF. DrugBank

<!--Datatype properties for the resource that corresponds to Lepirudin-->
http://bio2rdf.org/page/drugbank:DB00001 rdfs:label "Lepirudin"
http://bio2rdf.org/page/drugbank:DB00001 :absorption "Bioavailability is 100% following injection"

<!--Object property Dosage and route for Lepirudin-->
http://bio2rdf.org/pharmgkb_vocabulary:361ee98c3d82f85e8095179351912761 rdfs:label "Intravenous"

<!--Interaction between two drugs: Lepirudin and Ginkgo biloba-->
http://bio2rdf.org/drugbank:DB00374 rdfs:label "Ginkgo biloba"
http://bio2rdf.org/drugbank_resource:DB00001_DB00374 rdfs:label "DDI between Lepirudin and Treprostinil - The prostacyclin analogue, Treprostinil, increases the risk of bleeding when combined with the anticoagulant, Lepirudin. Monitor for increased bleeding during concomitant therapy."
Bio2RDF. DrugBank

<!--Datatype properties for the resource that corresponds to Lepirudin-->
http://bio2rdf.org/page/drugbank:DB00001 rdfs:label "Lepirudin"
http://bio2rdf.org/page/drugbank:DB00001:absorption "Bioavailability is 100% following injection"

<!--Object property Dosage and route for Lepirudin-->  
http://bio2rdf.org/pharmgkb_vocabulary:361ee98c3d82f85e8095179351912761 rdfs:label "Intravenous"

<!--Interaction between two drugs: Lepirudin and Ginkgo biloba-->  
http://bio2rdf.org/drugbank:DB00374 rdfs:label "Ginkgo biloba"
http://bio2rdf.org/drugbank_resource:DB00001_DB00374 rdfs:label "DDI between Lepirudin and Treprostinil - The prostacyclin analogue, Treprostinil, increases the risk of bleeding when combined with the anticoagulant, Lepirudin. Monitor for increased bleeding during concomitant therapy."
Bio2RDF. DrugBank

<!--Datatype properties for the resource that corresponds to Lepirudin-->
http://bio2rdf.org/page/drugbank:DB00001 rdfs:label "Lepirudin"
http://bio2rdf.org/page/drugbank:DB00001 :absorption "Bioavailability is 100% following injection"

http://bio2rdf.org/pharmgkb_vocabulary:361ee98c3d82f85e8095179351912761 rdfs:label "Intravenous"

http://bio2rdf.org/drugbank:DB00374 rdfs:label "Ginkgo biloba"
http://bio2rdf.org/drugbank_resource:DB00001_DB00374 rdfs:label "DDI between Lepirudin and Treprostinil - The prostacyclin analogue, Treprostinil, increases the risk of bleeding when combined with the anticoagulant, Lepirudin. Monitor for increased bleeding during concomitant therapy."
Example: What is the description and route of elimination for Lepirudin?

```
column | property | inSelect | searchFor
--- | --- | --- | ---
```

```
NP
  NP
    DT the
    NN description
  CC and
  NN route-of-elimination
  IN for
    NN np
      NNP Lepirudin
```
Example: What is the description and route of elimination for Lepirudin?

```
NP
   /    \
  NP    PP
    /     \
   DT  NN   CC  NN  IN  NP
      the  description  and  route-of-elimination  for
               NNP
                  Lepirudin

<table>
<thead>
<tr>
<th>column</th>
<th>property</th>
<th>inSelect</th>
<th>searchFor</th>
</tr>
</thead>
<tbody>
<tr>
<td>description</td>
<td>⟨<a href="http://purl.org/dc/terms/description%E2%9F%A9">http://purl.org/dc/terms/description⟩</a></td>
<td>true</td>
<td>null</td>
</tr>
</tbody>
</table>
```
Example: What is the description and route of elimination for Lepirudin?

<table>
<thead>
<tr>
<th>column</th>
<th>property</th>
<th>inSelect</th>
<th>searchFor</th>
</tr>
</thead>
<tbody>
<tr>
<td>description</td>
<td>⟨<a href="http://purl.org/dc/terms/description%E2%9F%A9">http://purl.org/dc/terms/description⟩</a></td>
<td>true</td>
<td>null</td>
</tr>
<tr>
<td>name</td>
<td>⟨<a href="http://www.w3.org/2000/01/rdf-schema#label%E2%9F%A9">http://www.w3.org/2000/01/rdf-schema#label⟩</a></td>
<td>false</td>
<td>Lepirudin</td>
</tr>
</tbody>
</table>
Example: What is the description and route of elimination for Lepirudin?

<table>
<thead>
<tr>
<th>column</th>
<th>property</th>
<th>inSelect</th>
<th>searchFor</th>
</tr>
</thead>
<tbody>
<tr>
<td>description</td>
<td>⟨<a href="http://purl.org/dc/terms/description%E2%9F%A9">http://purl.org/dc/terms/description⟩</a></td>
<td>true</td>
<td>null</td>
</tr>
<tr>
<td>name</td>
<td>⟨<a href="http://www.w3.org/2000/01/rdf-schema#label%E2%9F%A9">http://www.w3.org/2000/01/rdf-schema#label⟩</a></td>
<td>false</td>
<td>Lepirudin</td>
</tr>
</tbody>
</table>
Example: What is the description and route of elimination for Lepirudin?

```
SELECT ?description ?routeOfElimination
WHERE {

  FILTER REGEX ( str(?name), "Lepirudin", "i" ).}
```
Pattern 1

What is/are the list of properties for/of drug name?

or

Find the list of properties for drug name, where properties can be binary or n-ary with $n \geq 3$. 
Pattern 2 - properties of classes of drugs

- 2.1 inclusion or absence of inclusion into a certain category or presence of interaction: drugs that are (not) in Anticoagulant category
- 2.2 existence of synonymous relation: that are synonymous with Lepirudin
- 2.3 absence of interaction with another drug: that do not interact with Lepirudin
- 2.4 combination of the first three classes: that interact with Ginkgo Biloba, are in anticoagulant category, and do not interact with Lepirudin
Pattern 2 - subpatterns

anticoagulants category

are synonymous with Lepirudin

Anca Marginean si Oana Marc Towards Querying Bioinformatic Linked Data in Natural Language
Open questions

- What type of drug is Hirudin?
- What are the drugs from analgesics category?
- Find the drug categories that Lepirudin is part of.

Closed question - with yes/no answer

- Is there an interaction between Thrombin and Lepirudin?
- Is it bad/safe/dangerous to combine Lepirudin with Thrombin?
Conclusions

Our first results in querying a component of Linked Data - Bio2RDF Data, in natural language.

- the NL sentence is parsed to its structure tree
- the main entities are extracted in a pattern based manner
- the SPARQL query is built and executed

**PRO:** the lexicon and the structural patterns are built from the underlying vocabularies → portability in terms of ontology

**CONS:** BIO2RDF - source of ambiguity is strongly reduced.

Future work:

- to extend the set of patterns and the functions to compose them
- to improve the relation between lexicon and vocabulary
- to apply (statistical) learning machine (WASP) to refine the set of patterns