

# **Pokročilé dotazování jazykem SPARQL v databázích malých molekul**

**Jakub Galgonek**

Ústav organické chemie a biochemie AV ČR, v.v.i.  
skupina Bioinformatika  
Praha, Česká Republika

14. červen 2016

# Outline

---

- Sémantický web a linked-data technologie
  - Resource Description Framework
  - Web Ontology Language
  - SPARQL
- nedostatky informačních systémů založených na těchto technologiích
- náš přístup k řešení těchto nedostatků
- webová aplikace demonstrující naše řešení

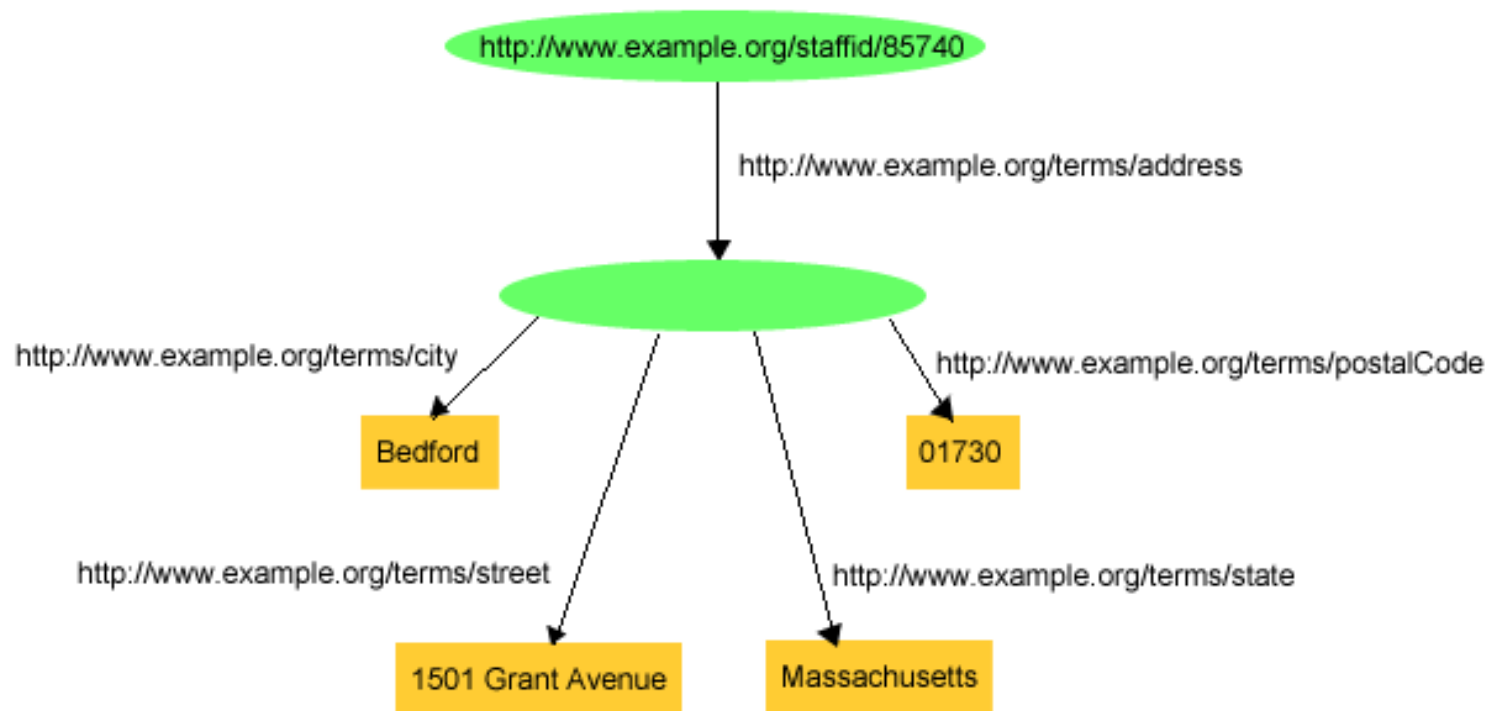
# Sémantický web

- Tim Berners-Lee (2001)
  - umožnit využívat web i programy
  - přidělit informacím význam
  - identifikace dat
    - globální identifikátory
  - konceptualizace dat
    - ontologie

# Resource Description Framework I.

- framework umožňující vyjádřit informace o zdrojích (resources)
- informace vyjadřuje jako jednoduchá tvrzení
  - triple: subject predicate object
  - subjekt má vlastnost predikate, jejíž hodnota je object
- typy zdrojů:
  - IRI: zdroje s explicitním jménem
  - blank nodes: zdroje bez explicitního jména
  - literály: základní hodnoty (řetězce, čísla, datumy, ...)

# Resource Description Framework II.



# Web Ontology Language

- jazyk pro popis ontologie dat
- popisuje třídy zdrojů
  - subclass hierarchie
- popisuje vlastnosti (properties)
  - doména
  - obor hodnot
  - subproperty hierarchii

# SPARQL query language

- dotazovací jazyk nad RDF daty
- základní koncept: triple vzor
  - stejný jako RDF triple, ale může obsahovat proměnné
- výsledkem vyhodnocení je mapování řešení
- vzory lze kombinovat různými jazykovými konstrukcemi

# Nedostatky

---

- není podporováno volání procedur
- psaní SPARQL dotazů je náchylné k chybám
- prezentace výsledků je buď uživatelsky nevlídná nebo neflexibilní

# SPARQL: Volání procedur I.

- není vhodné rozšiřovat syntax jazyka
- volání procedury reprezentováno tripletem
  - definovaný predikát určuje proceduru
  - objekt určuje sadu parametrů
    - jeho dílčí vlastnosti pak jednotlivé parametry
  - subjekt určuje výsledky
    - jeho dílčí vlastnosti pak jednotlivé složky výsledku

# SPARQL: Volání procedur II.

```
1.  DEFINE input:inference "ontology"
2.
3.  SELECT ?COMPOUND (max(?SCORE) as ?MAXSCORE) WHERE
4.  {
5.    ?ATB chebix:isA / chebi:identifier "17334".
6.    ?ATB chebix:molFile ?MOLFILE.
7.
8.    [ orchem:compound ?COMPOUND; orchem:score ?SCORE ]
9.      orchem:similaritySearch
10.        [ orchem:query ?MOLFILE;
11.          orchem:queryType "MOL";
12.          orchem:cutoff 0.9 ].
13.
14.  MINUS
15.  {
16.    ?COMPOUND chebix:hasRole / chebi:identifier "33281".
17.  }
18. }
19. GROUP BY ?COMPOUND
```

# SPARQL: Volání procedur III.

---

- překlad probíhá do SPARQL/SQL
- výhody našeho přístupu:
  - možnost definovat ontologii pro volání procedur
  - možnost definovat výchozí hodnoty parametrů
  - parametry lze uvádět v libovolném pořadí

# Sémantická kontrola I.

- dotazy jsou kontrolovány proti ontologii dat
- kontroly zahrnují:
  - existenci property IRI
  - správnost typu literálu
  - návaznost v property path
  - konzistenci použitých proměnných
    - kontroluje, zda typ proměnné odvozený ze specifické pozice v dotazu je konzistentní se zbytkem dotazu

# Konzistence proměnných I.

- dotaz je přeložen do syntaktického stromu
- strom je kontrolován od listů ke kořeni
- pro každý uzel syntaktického stromu:
  - class registr: celkové typy proměnných odvozené z daného podstromu
  - location registr: výskyty proměnných v daném podstromu a příslušný lokální typ
- group pattern uzel: informace uložená v location registru libovolného syna je kontrolována vůči class registrům jeho sourozenců

# Konzistence proměnných II.

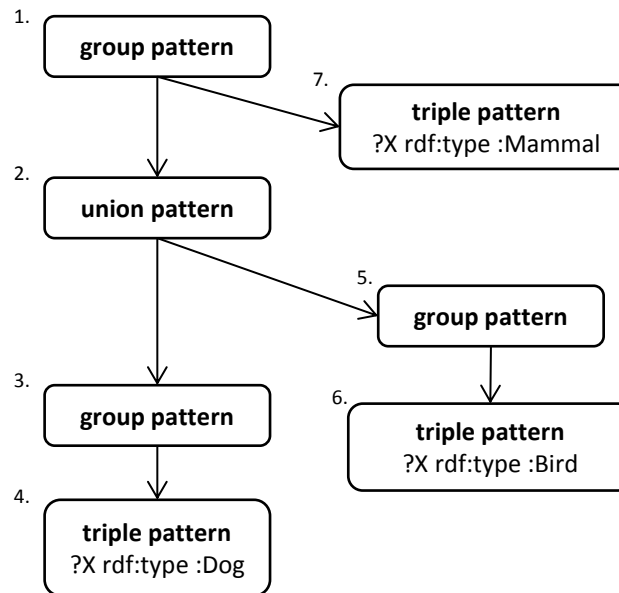
```
1. PREFIX : <http://ex.com>
2.
3. SELECT * WHERE
4. {
5.   {
6.     ?X rdf:type :Dog.
7.   }
8.   UNION
9.   {
10.    ?X rdf:type :Bird.
11.   }
12.   ?X rdf:type :Mammal.
13. }
```

**A**

# Konzistence proměnných II.

```
1. PREFIX : <http://ex.com>
2.
3. SELECT * WHERE
4. {
5.   {
6.     ?X rdf:type :Dog.
7.   }
8.   UNION
9.   {
10.    ?X rdf:type :Bird.
11.  }
12.  ?X rdf:type :Mammal.
13. }
```

**A**

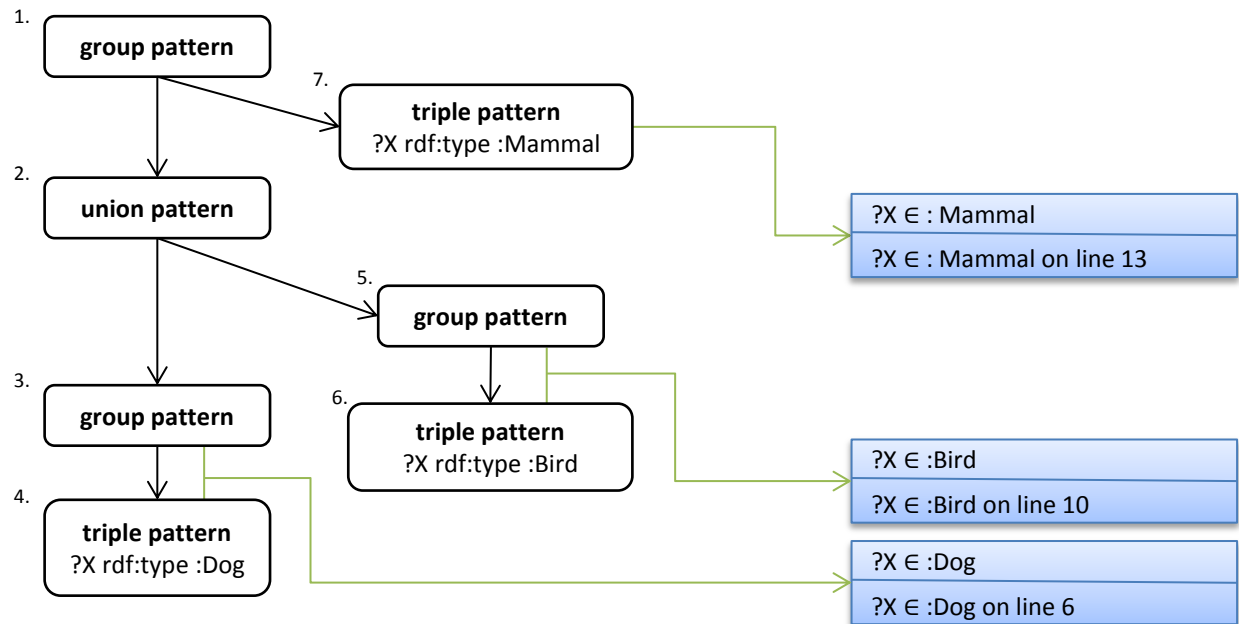


**B**

# Konzistence proměnných II.

```
1. PREFIX : <http://ex.com>
2.
3. SELECT * WHERE
4. {
5.   {
6.     ?X rdf:type :Dog.
7.   }
8.   UNION
9.   {
10.    ?X rdf:type :Bird.
11.  }
12.  ?X rdf:type :Mammal.
13. }
```

A



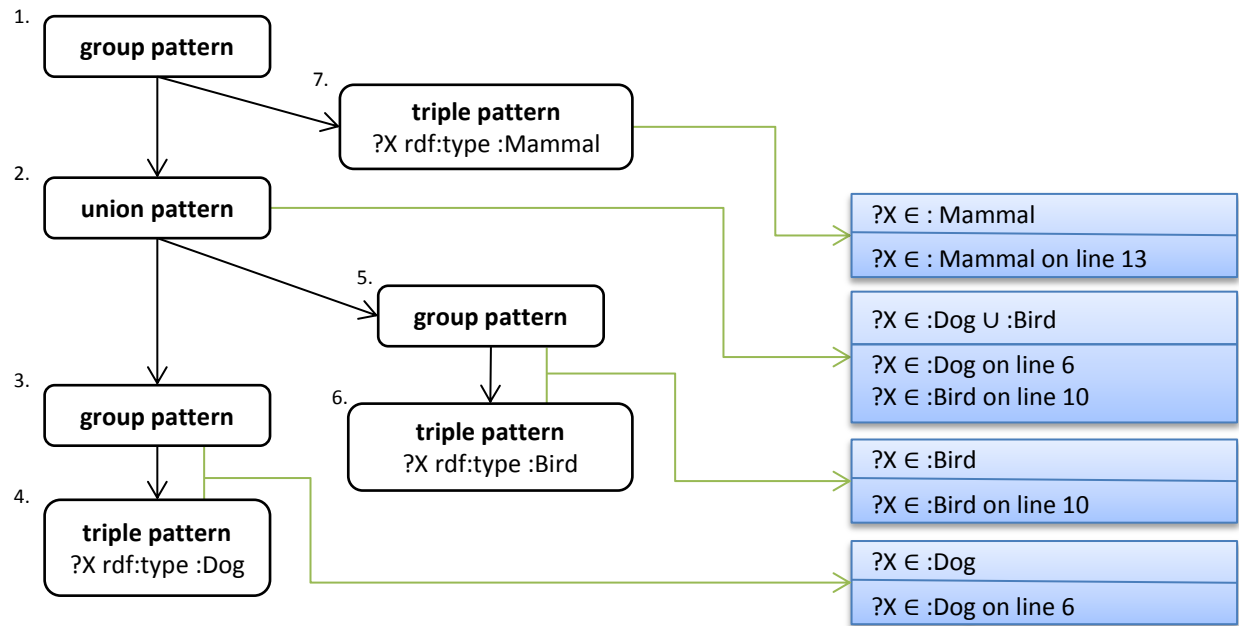
B

C

# Konzistence proměnných II.

```
1. PREFIX : <http://ex.com>
2.
3. SELECT * WHERE
4. {
5.   {
6.     ?X rdf:type :Dog.
7.   }
8.   UNION
9.   {
10.    ?X rdf:type :Bird.
11.  }
12.  ?X rdf:type :Mammal.
13. }
```

A



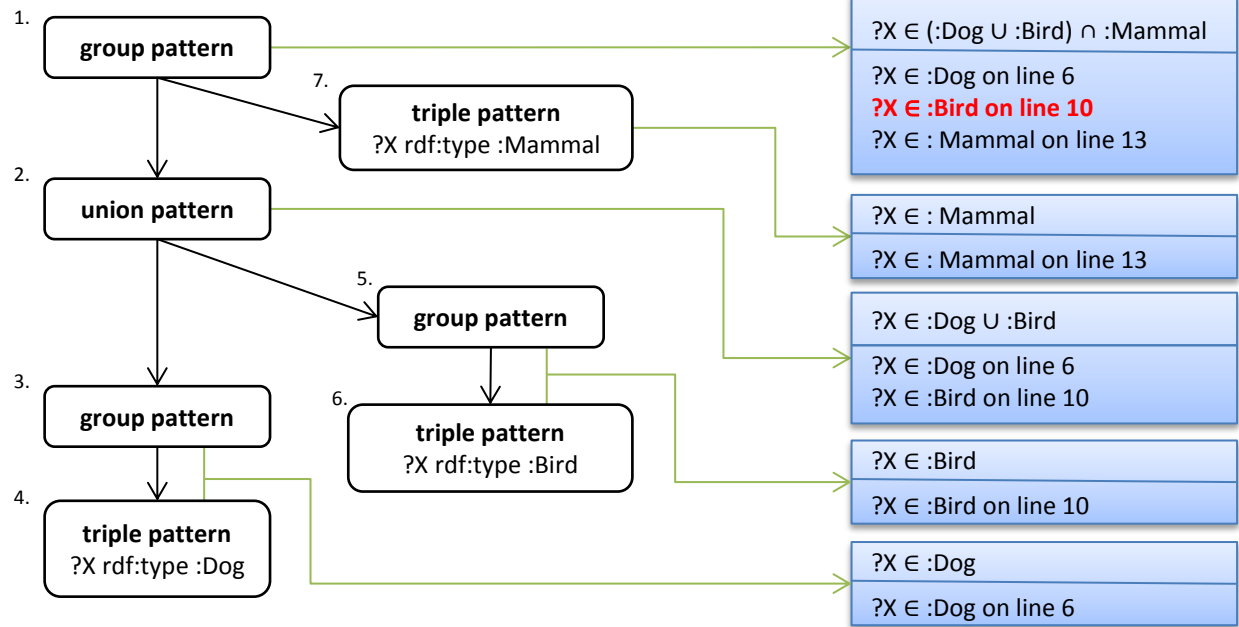
B

C

# Konzistence proměnných II.

```
1. PREFIX : <http://ex.com>
2.
3. SELECT * WHERE
4. {
5.   {
6.     ?X rdf:type :Dog.
7.   }
8.   UNION
9.   {
10.    ?X rdf:type :Bird.
11.  }
12.  ?X rdf:type :Mammal.
13. }
```

A



B

C

# Prezentace výsledků I.

- pro každou třídu: šablony určující způsob vizualizace instancí této třídy
- používáme Apache Velocity
  - velmi jednoduchý a rozšiřitelný
  - šablony obsahují:
    - prostý text: přímo okopírován na výstup
    - reference: reprezentují dynamický obsah
    - direktivy: kontrolují generování obsahu
      - definujeme několik vlastních directiv

# Prezentace výsledků II.

```
1.  #sparql($roles)
2.    SELECT ?ROLEREF ?NAME WHERE
3.    {
4.      $entity chebix:hasRole ?ROLEREF.
5.      ?ROLEREF chebix:chebiName ?NAME.
6.    }
7.  #end
8.
9.  <ul>
10.    #foreach($role in $roles)
11.      <li>
12.        #url($role.ROLEREF)
13.        #escapeHTML($role.NAME)
14.      #end
15.    </li>
16.  #end
17. </ul>
```

# Webová aplikace – chemRDF I.

- založená na ChEBI konvertované do RDF
- zahrnuje:
  - modifikovaný OrChem software umožňující podobnostní vyhledávání
  - ontologie pro sémantickou kontrolu
  - sadu šablon pro vizualizaci výsledků
- <http://bioinfo.uochb.cas.cz/projects/chemRDF>
- Advanced SPARQL querying in small molecule databases  
Galgonek et al. J Cheminform (2016) 8:31

# Webová aplikace – chemRDF II.

IOCB RDF Platform: ChEBI data (technical preview)

Query

Open Examples Save Run

```
1 DEFINE input:inference "ontology"
2
3 SELECT ?COMPOUND ?ATB ?SCORE WHERE
4 {
5   # ?ATB is a penicillin antibiotic
6   ?ATB chebi:isA / chebi:identifier
7   "17334".
8
9   # ?COMPOUND is similar to ?ATB structure
10  [ orchem:compound ?COMPOUND;
11    orchem:score ?SCORE ]
12    orchem:similaritySearch [
13      orchem:query [ ^chebi:molFile
14        ?ATB ];
15      orchem:queryType "MOL";
16      orchem:cutoff 0.9 ].
17
18   # ?COMPOUND has not the antimicrobial
19   agent role
20   MINUS
21   {
22     ?COMPOUND chebi:hasRole /
23     chebi:identifier "33281".
24   }
25 }
26 ORDER BY DESC (?SCORE)
```

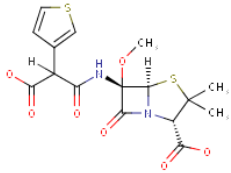
Result

any

| COMPOUND                                                                                                                                               | ATB                                                                                                                                    | SCORE      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|------------|
|  <b>CHEBI:52438</b><br>temocillin(2-)<br>Stars: ★★★★★                 |  <b>CHEBI:51817</b><br>temocillin<br>Stars: ★★★★★     | 1.0        |
|  <b>CHEBI:31615</b><br>flu-cloxacillin sodium<br>Stars: ★★★★★         |  <b>CHEBI:5098</b><br>flu-cloxacillin<br>Stars: ★★★★★ | 0.9823009  |
|  <b>CHEBI:51697</b><br>apalcillin sodium<br>Stars: ★★★★★              |  <b>CHEBI:51691</b><br>apalcillin<br>Stars: ★★★★★     | 0.981982   |
|  <b>CHEBI:51700</b><br>apalcillin potas-sium<br>Stars: ★★★★★          |  <b>CHEBI:51691</b><br>apalcillin<br>Stars: ★★★★★     | 0.981982   |
|  <b>CHEBI:34691</b> di-cloxacillin sodium<br>Stars: ★★★★★           |  <b>CHEBI:4511</b> di-cloxacillin<br>Stars: ★★★★★   | 0.98165137 |
|  <b>CHEBI:52067</b> me-zlocillin sodium<br>Stars: ★★★★★             |  <b>CHEBI:6919</b> me-zlocillin<br>Stars: ★★★★★     | 0.98165137 |
|  <b>CHEBI:52068</b> me-zlocillin sodium monohydrate<br>Stars: ★★★★★ |  <b>CHEBI:6919</b> me-zlocillin<br>Stars: ★★★★★     | 0.98165137 |

Details Properties About

chebiID:52438



ChEBI Name

**temocillin(2-)**

ChEBI ID

**CHEBI:52438**

Stars

★★★★★

|              |                                                                                                                                                                               |
|--------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Formula      | C16H16N2O7S2                                                                                                                                                                  |
| Net Charge   | -2                                                                                                                                                                            |
| Average Mass | 412.437                                                                                                                                                                       |
| InChI        | InChI=1S/C16H18N2O7S2/c1-15(2)9(12(22)23)18-13(24)16(25-3,14(18)27-15)17-10(19)8(11(20)21)7-4-5-26-6-7/h4-6,8-9,14H,1-3H3,(H,17,19)(H,20,21)(H,22,23)/p-2/t8?9-,14+,16-/m0/s1 |
| InChIKey     | BVCKFLJARNKCSS-DWPRYXJFSA-L                                                                                                                                                   |
| SMILES       | [H][C@]12SC(C)(C)[C@@H](N1C(=O)[C@]2(NC(=O)C(H)(C([O-])=O)c1ccsc1)OC)C([O-])=O                                                                                                |

ChEBI Ontology

Outgoing

Incoming

temocillin(2-) (CHEBI:52438) is a dicarboxylic acid dianion (CHEBI:28965)

temocillin(2-) (CHEBI:52438) is conjugate base of temocillin (CHEBI:51817)

temocillin disodium (CHEBI:52434) has part temocillin(2-) (CHEBI:52438)

temocillin (CHEBI:51817) is conjugate acid of temocillin(2-) (CHEBI:52438)

# Výhled do budoucna

- integrace s dalších zdrojů dat
  - PubChem
- soustředit se na konkrétní potřeby
- možné změny paradigmatu
  - relační databáze mapovaná do RDF
- rozšíření týmu
  - zájemci vítáni
    - diplomky, postdoci, ...
  - kontakt: [galgonek@uochb.cas.cz](mailto:galgonek@uochb.cas.cz)

# Poděkování



studenti:

- Tomáš Hurt
- Vendula Michlíková
- Petr Onderka
- Jan Schwarz

# Konec

---

**Děkuji za pozornost!**