Querying Semantic Web Data

Lalana Kagal
Decentralized Information Group
MIT CSAIL

Eric Prud'hommeaux
Sanitation Engineer
World Wide Web Consortium
SPARQL Program

- Graph patterns
- Motivations for RDF
- RDF description
- Turtle
- Motivations for SPARQL
- SPARQL
  - Structure
  - Mechanics
  - Resources
Graph patterns

Presenting three languages that use graph patterns:
- SPARQL
- N3
- AIR
Motivations for RDF

- Simple, consistent data model
- Uses web architecture for web scalability
- Glamorous use cases

Image courtesy http://clip.dia.fi.upm.es/~logalg/slides/
RDF for Drug Discovery

- Using the Semantic Web: Precise Answers to Complex Questions:
  - Find me genes involved in signal transduction that are related to pyramidal neurons.
Google

pyramidal neurons signal transduction

Web Books

Results 1 - 10 of about 223,000 for pyramidal neurons signal transduction. (0)

Book results for pyramidal neurons signal transduction

Cerebral Signal Transduction - by Maarten Eduard Anton Reith - 440 pages
Neuroprotective Signal Transduction - by Mark Paul. Mattson - 347 pages
Toxins And Signal Transduction - by Yehuda Gutman, Philip Lazarovici - 520 pages

Neurotrophin-3 and brain-derived neurotrophic factor activate ...
... and brain-derived neurotrophic factor activate multiple signal transduction events but are not survival factors for hippocampal pyramidal neurons. ...
www.ihop-net.org/UniPub/iHop/pm/646092.html?pmid=8752100 - 12k -
Cached - Similar pages - Note this

K+ channel regulation of signal propagation in dendrites of ...
Pyramidal neurons receive tens of thousands of synaptic inputs on their dendrites. ...
Signal Transduction* Substances Potassium Channel Blockers ...

Dopamine modulates inwardly rectifying potassium currents in ...
Using outside-out patches of mPFC pyramidal neurons, which preclude involvement of ...
Signal Transduction/drug effects Signal Transduction/physiology ...

Loss of Hippocampal CA3 Pyramidal Neurons in Mice Lacking STAM1 ...
Loss of Hippocampal CA3 Pyramidal Neurons in Mice Lacking STAM1 ... and to be involved in the regulation of intracellular signal transduction mediated by ...
mcb.asm.org/cgi/content/abstract/21/11/3807 - Similar pages - Note this
1: Naimark A, Barkai E, Matar MA, Kaplan Z, Kozlovsky N, Cohen H.

Upregulation of neurotrophic factors selectively in frontal cortex in response to olfactory discrimination learning.

Neural Plast. 2007::13427.
PMID: 17710248 [PubMed - in process]


The blockade of K(+)--ATP channels has neuroprotective effects in an in vitro model of brain ischemia.

PMID: 17678973 [PubMed - indexed for MEDLINE]

3: Schmidt-Hieber C, Jonas P, Bischofberger J.

Subthreshold dendritic signal processing and coincidence detection in dentate gyrus granule cells.

PMID: 17670990 [PubMed - indexed for MEDLINE]

4: Alvarez VA, Ridenour DA, Sabatini BL.

Distinct structural and ionotropic roles of NMDA receptors in controlling spine and synapse stability.

PMID: 17626197 [PubMed - indexed for MEDLINE]

5: Smith SS, Gong QH.

Ethanol effects on GABA-activated current in a model of increased alpha/beta/delta GABA A receptor.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2580</td>
<td>PubMed: biomedical literature citations and abstracts</td>
</tr>
<tr>
<td>959</td>
<td>PubMed Central: free, full text journal articles</td>
</tr>
<tr>
<td>none</td>
<td>Site Search: NCBI web and FTP sites</td>
</tr>
<tr>
<td>2</td>
<td>Books: online books</td>
</tr>
<tr>
<td>4</td>
<td>OMIM: online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>none</td>
<td>OMIA: online Mendelian Inheritance in Animals</td>
</tr>
<tr>
<td>10</td>
<td>CoreNucleotide: Core subset of nucleotide sequence records</td>
</tr>
<tr>
<td>none</td>
<td>EST: Expressed Sequence Tag records</td>
</tr>
<tr>
<td>34</td>
<td>GSS: Genome Survey Sequence records</td>
</tr>
<tr>
<td>16</td>
<td>Protein: sequence database</td>
</tr>
<tr>
<td>5</td>
<td>Genome: whole genome sequences</td>
</tr>
<tr>
<td>none</td>
<td>Structure: three-dimensional macromolecular structures</td>
</tr>
<tr>
<td>none</td>
<td>Taxonomy: organisms in GenBank</td>
</tr>
<tr>
<td>none</td>
<td>SNP: single nucleotide polymorphism</td>
</tr>
<tr>
<td>35</td>
<td>Gene: gene-centered information</td>
</tr>
<tr>
<td>10</td>
<td>HomoloGene: eukaryotic homology groups</td>
</tr>
<tr>
<td>none</td>
<td>PubChem Compound: unique small molecule chemical structures</td>
</tr>
<tr>
<td>none</td>
<td>PubChem Substance: deposited chemical substance records</td>
</tr>
<tr>
<td>none</td>
<td>Genome Project: genome project information</td>
</tr>
<tr>
<td>2</td>
<td>dbGaP: genotype and phenotype</td>
</tr>
<tr>
<td>none</td>
<td>UniGene: gene-oriented clusters of transcript sequences</td>
</tr>
<tr>
<td>none</td>
<td>CDD: conserved protein domain database</td>
</tr>
<tr>
<td>none</td>
<td>3D Domains: domains from Entrez Structure</td>
</tr>
<tr>
<td>none</td>
<td>UniSTS: markers and mapping data</td>
</tr>
<tr>
<td>none</td>
<td>PopSet: population study data sets</td>
</tr>
<tr>
<td>none</td>
<td>GEO Profiles: expression and molecular abundance profiles</td>
</tr>
<tr>
<td>none</td>
<td>GEO DataSets: experimental sets of GEO data</td>
</tr>
<tr>
<td>none</td>
<td>Cancer Chromosomes: cytogenetic databases</td>
</tr>
<tr>
<td>none</td>
<td>PubChem BioAssay: bioactivity screens of chemical substances</td>
</tr>
<tr>
<td>none</td>
<td>GENSAT: gene expression atlas of mouse central nervous system</td>
</tr>
<tr>
<td>none</td>
<td>Probe: sequence-specific reagents</td>
</tr>
<tr>
<td>1394</td>
<td>Protein Clusters: a collection of related protein sequences</td>
</tr>
</tbody>
</table>
Integrate databases ...

- MeSH
- PubMed
- Entrez Gene
- Gene Ontology
- ...

Diagram showing various databases connected, including:
- Gene Ontology
- Reactome
- PDSPki
- NeuronDB
- BAMS
- BrainPharm
- MESH
- PubChem
- SWAN
- AlzGene
- Homologene
- Mammalian Phenotype
- Allen Brain Atlas
so that one query 

Mesh: Pyramidal Neurons 

Pubmed: Journal Articles 

Entrez Gene: Genes 

GO: Signal Transduction
... spans several DBs ...
... to yield cross-specialty information

<table>
<thead>
<tr>
<th>Gene/Protein</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRD1, 1812</td>
<td>adenylate cyclase activation</td>
</tr>
<tr>
<td>ADRB2, 154</td>
<td>adenylate cyclase activation</td>
</tr>
<tr>
<td>ADRB2, 154</td>
<td>arrestin mediated desensitization of G-protein coupled receptor protein signaling pathway</td>
</tr>
<tr>
<td>DRD1IP, 50632</td>
<td>dopamine receptor signaling pathway</td>
</tr>
<tr>
<td>DRD1, 1812</td>
<td>dopamine receptor, adenylate cyclase activating pathway</td>
</tr>
<tr>
<td>DRD2, 1813</td>
<td>dopamine receptor, adenylate cyclase inhibiting pathway</td>
</tr>
<tr>
<td>GRM7, 2917</td>
<td>G-protein coupled receptor protein signaling pathway</td>
</tr>
<tr>
<td>GNG3, 2785</td>
<td>G-protein coupled receptor protein signaling pathway</td>
</tr>
<tr>
<td>GNG12, 55970</td>
<td>G-protein coupled receptor protein signaling pathway</td>
</tr>
<tr>
<td>DRD2, 1813</td>
<td>G-protein coupled receptor protein signaling pathway</td>
</tr>
<tr>
<td>ADRB2, 154</td>
<td>G-protein coupled receptor protein signaling pathway</td>
</tr>
<tr>
<td>CALM3, 808</td>
<td>G-protein coupled receptor protein signaling pathway</td>
</tr>
<tr>
<td>HTR2A, 3356</td>
<td>glutamate signaling pathway</td>
</tr>
<tr>
<td>TRD1, 1812</td>
<td>G-protein signaling, coupled to cyclic nucleotide second messenger</td>
</tr>
<tr>
<td>SSTR5, 6755</td>
<td>G-protein signaling, coupled to cyclic nucleotide second messenger</td>
</tr>
<tr>
<td>MTNR1A, 4543</td>
<td>G-protein signaling, coupled to cyclic nucleotide second messenger</td>
</tr>
<tr>
<td>CNR2, 1269</td>
<td>G-protein signaling, coupled to cyclic nucleotide second messenger</td>
</tr>
<tr>
<td>HTR6, 3362</td>
<td>G-protein signaling, coupled to cyclic nucleotide second messenger</td>
</tr>
<tr>
<td>GRIK2, 2898</td>
<td>glutamate signaling pathway</td>
</tr>
<tr>
<td>GRIN1, 2902</td>
<td>glutamate signaling pathway</td>
</tr>
<tr>
<td>GRIN2A, 2903</td>
<td>glutamate signaling pathway</td>
</tr>
<tr>
<td>GRIN2B, 2904</td>
<td>glutamate signaling pathway</td>
</tr>
<tr>
<td>ADAM10, 102</td>
<td>integrin mediated signaling pathway</td>
</tr>
<tr>
<td>GRM7, 2917</td>
<td>negative regulation of adenylate cyclase activity</td>
</tr>
<tr>
<td>LRP1, 4035</td>
<td>negative regulation of Wnt receptor signaling pathway</td>
</tr>
<tr>
<td>ADAM10, 102</td>
<td>Notch receptor processing</td>
</tr>
<tr>
<td>ASCL1, 429</td>
<td>Notch signaling pathway</td>
</tr>
<tr>
<td>HTR2A, 3356</td>
<td>serotonin receptor signaling pathway</td>
</tr>
<tr>
<td>ADRB2, 154</td>
<td>transmembrane receptor protein tyrosine kinase activation (dimerization)</td>
</tr>
<tr>
<td>PTPRC, 5793</td>
<td>transmembrane receptor protein tyrosine kinase signaling pathway</td>
</tr>
<tr>
<td>EPHA4, 2043</td>
<td>transmembrane receptor protein tyrosine kinase signaling pathway</td>
</tr>
<tr>
<td>NRTN, 4902</td>
<td>transmembrane receptor protein tyrosine kinase signaling pathway</td>
</tr>
</tbody>
</table>
Patient data

- Patient identifier
- Medical history
- Family medical history
- Health-related behaviour

```xml
<?xml version="1.0"?>
<ClinicalDocument transformation="hl7-rim-to-pomr.xslt">
  <recordTarget>
    <patientRole>
      <patientPatient>
        <name>
          <given>Henry</given>
          <family>Levin</family>
        </name>
        <administrativeGender code="M"/>
        <birthTime value="19320924"/>
      </patientPatient>
    </patientRole>
  </recordTarget>
...
RDF works for all of this ...
... regardless of its source
What does RDF provide?

- Common (simple) model to for all this data.
- Incentive and infrastructure to re-use terms when possible and invent terms when necessary.
- Simple and complex ontology languages (RDFS and OWL).
- Intuitive re-use of now-familiar web topology.
- Scalable — partial (monotonic) reasoning allowed. Apps need not be re-written for each extension to a document.
How do we write RDF?

- Name resources and relationships with URIs
  - e.g. http://people.apache.org/~oshani/foaf.rdf#me represents a person
- Express statements as subject, predicate, object
- Write the triples in
  - RDF/XML: Standard serialization in XML
    <Description about="subject"><property>value</property></Description>
  - NTriples: Simple (verbose) reference serialization (for specifications only)
    <http://...subject> <http://...predicate> "value" .
  - N3 and Turtle: Developer-friendly serializations
    :subject :property "value" .
Turtle Introduction

- RDF triples analogous to one 3-place holds(s, p, o) predicate
  - oshani:me foaf:knows lalana:me .
  - holds(oshani:me foaf:knows lalana:me)
- Triples made from standard RDF terms:
  - IRIs: `<http://people.apache.org/~oshani/foaf.rdf#me>`, `<#me>`, oshani:me
  - Literals: "Oshani Seneviratne"
    - Typed literals: “3.14”^^xsd:float
    - Literals with language tags: “日本語”@ja
- Blank nodes: [], _:bob
  - Literal terms
• URI terms can be abbreviated using namespaces
  @prefix oshani: <http://people.apache.org/~oshani/foaf.rdf#> .
  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
  @prefix foaf: <http://xmlns.com/foaf/0.1/>
  oshani:me rdf:type foaf:Person .
• 'a' = <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  oshani:me rdf:type foaf:Person .
• In-line blank nodes
  oshani:me foaf:knows [ foaf:name "Lalana Kagal" ] .
Convenience Syntax

- Abbreviating repeated subjects:
  
  oshani:me rdf:type foaf:Person .
  oshani:me foaf:knows lalana:me .
  ...
  is the same as ...
  oshani:me rdf:type foaf:Person ; foaf:knows lalana:me .

- Abbreviating repeated subject/predicate pairs:
  
  oshani:me foaf:knows lalana:me .
  oshani:me foaf:knows timbl:i .
  ...
  is the same as ...
  oshani:me foaf:knows lalana:me , timbl:i .
Convenience Syntax

... is more succinctly represented as:

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://people.apache.org/~oshani/foaf.rdf#me>  a                foaf:Person ;
                                                   foaf:name        "Oshani Seneviratne" ;
Patient data in RDF

```xml
<?xml version="1.0"?>
<ClinicalDocument transformation="hl7-rim-to-pomr.xslt">
  <recordTarget>
    <patientRole>
      <patientPatient>
        <name>
          <given>Henry</given>
          <family>Levin</family>
        </name>
        <administrativeGender code="M"/>
        <birthTime value="19320924"/>
      </patientPatient>
      <screeningBP>
        <systolic>
          <AbsoluteMeasurement>
            <ex:unit>mm[Hg]</ex:unit>
            <r:value>132</r:value>
            <skos:prefLabel>Systolic BP</skos:prefLabel>
          </AbsoluteMeasurement>
        </systolic>
        <diastolic>
          <AbsoluteMeasurement>
            <ex:unit>mm[Hg]</ex:unit>
            ...  
          </AbsoluteMeasurement>
        </diastolic>
      </screeningBP>
    </patientRole>
  </recordTarget>
</ClinicalDocument>
```
RDF Resources

- RDF at the W3C - primer and specifications
- Semantic Web tools - community maintained list; includes triple store, programming environments, tool sets, and more
- 302 Semantic Web Videos and Podcasts - includes a section specifically on RDF videos
- RDF/XML sample patient data - complex model used in this tutorial
- Turtle sample patient data - complex model used in this tutorial
- Turtle simplified sample patient data - simple model used in this tutorial
Why SPARQL?

SPARQL is the query language of the Semantic Web. It lets us:

• Pull values from structured and semi-structured data
• Explore data by querying unknown relationships
• Perform complex joins of disparate databases in a single, simple query
• Transform RDF data from one vocabulary to another
SELECTing variables

• SPARQL variables bind to RDF terms
  – Ex. ?journal, ?disease, ?price

• Like SQL, we pick the variables we want from a query with a SELECT clause
  – Ex. SELECT ?article ?author ?published

• A SELECT query results in a table of values:

<table>
<thead>
<tr>
<th>?artist</th>
<th>?album</th>
<th>?times_platinum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Jackson</td>
<td>Thriller</td>
<td>27</td>
</tr>
<tr>
<td>Led Zeppelin</td>
<td>Led Zeppelin IV</td>
<td>22</td>
</tr>
<tr>
<td>Pink Floyd</td>
<td>The Wall</td>
<td>22</td>
</tr>
</tbody>
</table>
A triple pattern is an RDF triple that can have variables in any of the subject, predicate, or object positions.

Examples:

- Find countries and their capital cities:

- Given a FOAF URI, find the person's name:

- What direct relationships exist between two employees?
We can combine more than one triple pattern to retrieve multiple values and easily traverse an RDF graph:

- Find countries, their capital cities, and their populations:
  \[
  \text{?country geo:capital ?capital ; geo:population ?population .}
  \]

- Given a FOAF URI, find the person's name and friends' names:
  \[
  \langle \text{http://people.apache.org/~oshani/foaf.rdf#me} \rangle \text{ foaf:name ?name ; foaf:knows ?friend .}
  \]

  \[
  \text{?friend foaf:name ?friend_name .}
  \]

- Retrieve all third-line managers in the company:
  \[
  \text{?emp hr:managedBy ?first_line . ?first_line hr:managedBy ?second_line . ?second_line hr:managedBy ?third_line .}
  \]
SPARQL lets us query different RDF graphs in a single query. Consider movie reviews:

- Target one authoritative data source (What does Roger Ebert say?):
  
  ```sparql
  GRAPH <http://example.org/reviews/rogerebert> {
    ex:atonement rev:hasReview ?review .
  }
  ```
SPARQL lets us query different RDF graphs in a single query. Consider movie reviews:

- Relate multiple sources (How do my reviews compare to Ebert's?):
  - GRAPH <http://example.org/reviews/rogerebert> { 
  }
  GRAPH <http://example.org/reviews/me> { 
  }
GRAPH constraints

SPARQL lets us query different RDF graphs in a single query. Consider movie reviews:

- reviewers have given out perfect ratings?:
  
  GRAPH ?reviewer_graph {
  }
Besides selecting tables of values, SPARQL allows three other types of queries:

- **ASK** - returns a boolean answering, does the query have any results?
- **CONSTRUCT** - uses variable bindings to return new RDF triples
- **DESCRIBE** - returns server-determined RDF about the queried resources

SELECT and ASK results can be returned as XML or JSON. CONSTRUCT and DESCRIBE results can be returned via any RDF serialization (e.g. RDF/XML or Turtle).
Protocol Mechanics

The **SPARQL Protocol** is a simple method for asking and answering SPARQL queries over HTTP. A SPARQL URL is built from three parts:

1. The URL of a **SPARQL endpoint** (e.g. http://example.org/sparql)
2. (Optional, as part of the query string) The **graphs** to be queried against (e.g. named-graph-uri=http://example.org/reviews/ebert)
3. (As part of the query string) The **query** itself (e.g. query=SELECT...)

```plaintext
http://example.org/sparql?named-graph-uri=http%3A%2F%2Fexample.orgm%2Freviews%2Febert&query=SELECT+%3Freview_graph+WHERE+%7B%0D++GRAPH+%3Freview_graph+%7B%0D+++++%3Freview+rev%3Arating+10+.%0D%0A++%7D%0A+++++%3Freview+rev%3Arating+10+.%0D%0A%7D
```
Example Query: Henry Levin's Blood Pressure

```sparql
PREFIX r: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX snomed: <http://termhost.example/SNOMED/>

SELECT ?date ?sys ?dias ?position {
  ?p r:type galen:Patient ;  
    foaf:family_name "Levin" ;  
    foaf:firstName "Henry" .
  ?c edns:patient ?p ;  
    edns:screeningBP ?scr .
  ?scr dc:date ?date ;  
    edns:systolic [ r:value ?sys ] ;  
    edns:diastolic [ r:value ?dias ] ;  
    edns:posture ?position .
} ORDER by ?date
```

- data
- query
Example Query: Henry Levin's Blood Pressure

PREFIX r: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX snomed: <http://termhost.example/SNOMED/>

SELECT ?date ?sys ?dias {
  ?p r:type galen:Patient;
    foaf:family_name "Levin";
    foaf:firstName "Henry".
  ?c edns:patient ?p;
    edns:screeningBP ?scr .
  ?scr dc:date ?date ;
    edns:systolic [ r:value ?sys ];
    edns:diastolic [ r:value ?dias ];
    edns:posture snomed:_163035008 . # SNOMED:sitting
} ORDER by ?date

- data
- query
SPARQL Resources

- SPARQL Frequently Asked Questions
- SPARQL implementations - community maintained list of open-source and commercial SPARQL engines
- Public SPARQL endpoints - community maintained list
- SPARQL extensions - collection of SPARQL extensions implemented in various SPARQL engines