Graph-based queries of Semantic-Web integrated biological data

Marco Moretto[1,2], Alessandro Cestaro[2], Enrico Blanzieri[1] and Riccardo Velasco[2]

[1]University of Trento Via Sommarive, 14 38100 Trento, Italy

[2]Fondazione Edmund Mach Via Edmund Mach, 1 38010 S. Michele all'Adige, Trento , Italy

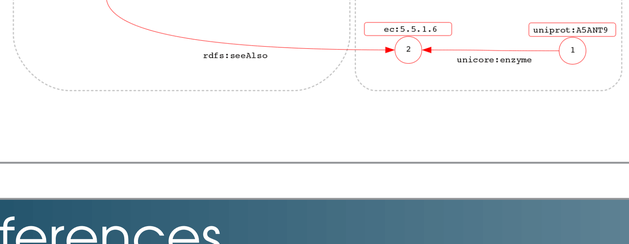
marco.moretto@iasma.it

Abstract

In the post-genomic era, life science researchers are faced with the need to manage and inspect a growing abundance of data and information. Data from different sources, both public and proprietary, have the most value when considered in the context of each other as they give more information. In order to answer questions that spans multiple fields in the biology domain without an integrated approach, a biologist needs to visit all data sources related to the problem and find relevant data. In the last years we have become witnesses of a growing interest for the Semantic Web technologies to integrate and query biological data. Semantic Web technologies were designed to meet the challenges of reduce the complexity of combining data from multiple sources, resolve the lack of widely accepted standards and manage highly distributed and mutable resources. However, Semantic Web standard technologies do not provide any tools to query integrated knowledge bases from a graph perspective, that is defining graph traversal patterns. For example, it is not possible to ask a query like "are enzyme A and compound B related?" without knowing the complete structure of the knowledge base. After exploring different alternatives we come up with the use of a graph traversal programming language on top of a triplestore in order to perform several path traversal queries on an integrated graph. We tested the feasibility of the approach integrating Uniprot, Gene Ontology, Chebi and Kegg resources posing queries of different complexity.

Gremlin

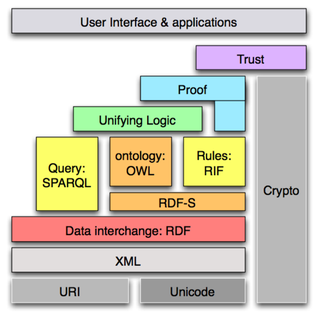
Is a domain specific programming language for graphs based on Groovy. It is not tied to a particular graph backend and its syntax allows for the representation of graph traversal expression succinctly



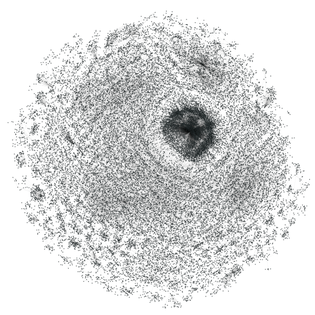
Example query: graph

Query *Given an enzyme and a compound, are they related?*

Overview

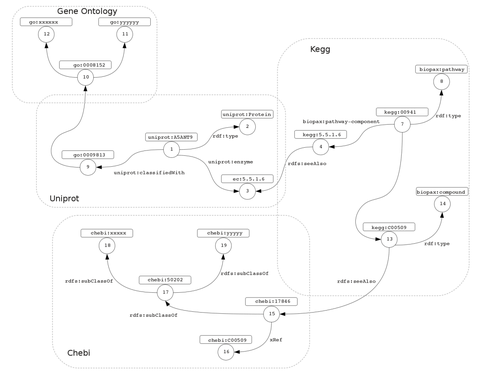


The Resource Description Framework (RDF) data model is based upon the idea of making statements about resources in the form of subject-predicate-object expressions, known as triples. From a database perspective, RDF can be considered an extension of graph database models.



Proposed solution

- Gathering resources from public repositories
- If necessary convert them into RDF format
- Store them into a Sesame triplestore
- Integrate them providing linking triples
- Query the integrated RDF graph using Gremlin



Example query: Gremlin code

Query *Given an enzyme and a compound, are they related?*

```
interactions= [g.v(g.uri('bpx:control')),g.v(g.uri('bpx:biochemicalReaction'))]
compound = g.v(g.uri('chebi:CHEBI_16077'))
enzyme = g.v(g.uri('uniprot:D7SXJ4'))
enzyme.outE[[label:g.uri('unicore:enzyme')]].inV.inE[[
 label:g.uri('rdfs:seeAlso')]].outV.inE.outV.loop(2){!interactions.contains((
  it.object.outE[[label:g.uri('rdf:type')]].inV >> 1))}.outE.inV.loop(2)
   {it.object != compound }
```

Conclusions and results

KB	Uniprot	GO	Chebi	KEGG
kb1	*Vitis vinifera*	only ids	only ids	*Vitis vinifera*
kb2	Eudicotyledons	only ids	only ids	*Vitis vinifera*
kb3	*Viridiplantae*	full	full	*Vitis vinifera*

KB	N of vertexes	N of edges	Loading time	Disk space
kb1	297.207	3.136.164	3 minutes	214 Mb
kb2	1.375.608	21.196.845	15 minutes	1.8 Gb
kb3	13.149.000	181.693.000	3 hours	15 Gb

- An integrated approach allows biologists to query different information resources without the need to visit all of them in order to find relevant data
- DBMS knowledge bases must be designed and modified with an idea of the type of queries they are going to answer
- Semantic Web technologies provide standard tools and technologies to easily integrate data from different sources
- SPARQL does not allow path traversal queries
- Graph-based approach allows to express queries like "are entity A and entity B related?"

References

[1] Goble, C. and Stevens, R. State of the nation in data integration for bioinformatics *Journal of biomedical informatics*, 41(5):687–693, 2008.

[2] Angles, R. and Gutierrez, C. Querying RDF data from a graph database perspective *The Semantic Web: Research and Applications, Springer*, 346–360 2005.

[3] Marko Rodriguez Gremlin *https://github.com/tinkerpop/gremlin/wiki*