Developing a New Analysis Framework - A Semantics Perspective

Evan Patton (pattoe@cs.rpi.edu), Spencer Norris (norris@rpi.edu)
Matthew Poegel (goosemt@rpi.edu), Deborah L. McGuinness (dim@cs.rpi.edu), Kristin Bennett (bennett@rpi.edu)
Rensselaer Polytechnic Institute / 110 8th Street / Troy, NY, 12180 USA
http://tw.rpi.edu

Introduction

The domain of bioinformatics has benefitted strongly from the integration of semantics in many large-scale systems. However, a persistent issue has been a lack of strong universal standards in how the data is defined and maintained across projects. While many attempts have been made to unify the way this information is represented, such as the Gene Ontology [1], many institutions continue to use internally-defined and highly lab-specific systems [2]. Because of this, it becomes difficult to tie the results of different datasets together, posing large roadblocks in designing services that rely on more than one database. The Semantic Numeric Exploration Technology (SemNExT) project is attempting to create a framework in which it becomes simple and intuitive to compose results irrespective of their origin or internal structure. This is accomplished by tying an internal ontology to the items provided by various data sources, which in turn are passed between the various sub-services, using JSON-LD as an intermediary. This data is then reasoned over and recombined using R scripts, the results of which are composed and returned to the user through an API. With this project, we hope to create a simplistic and reliable framework for composing linked data results irrespective of the domain focus.

Relating Different Data Sources

- Standardized gene identification conventions, such as the Entrez ID system developed by the National Center for Biotechnology Information [3], have made it easier to consistently reference the same gene across different data sources.
- If the encoding of this information for a particular data source is known, it can be accessed and cross-referenced in order to build out a dataset for that gene; accordingly, the data garnered from these results can be used to perpetuate the expansion of the knowledge base further down the chain.
- For example, a gene symbol is tied to the API and is referenced against a local instance of UMLS [4] to acquire a list of other aliases for the gene. These alternative labels are then used to find a matching URI in Uniprot [5], which is queried for in ReDrugS [6] in order to gather protein-protein and disease-gene interactions, as well as their respective provenance.

From Bioinformatics to Generalization

- The current iteration of the framework focuses exclusively on bioinformatics systems as a motivating example. Our data visualizations are tailored to this domain, demonstrating the results of the gene clustering analyses (see image on right).
- We hope to produce a system that can be easily generalized to any type of analysis application. This will be supplemented by the incorporation of a rules engine, which will enable the developer to define actions based upon the classes and datatypes of the results from different sets.

Underlying Technologies and Data Sources

- Datasources
  - SPARQL Endpoints/Triple Stores
    - Bio2RDF
    - Uniprot
    - ReDrugS
  - Relational Databases
    - StringDB
    - Ensembl
  - Unified Medical Language System (UMLS)
  - JSON-LD
- R Scripting Language
- CherryPy (Server-Side Python)
- Docker, Docker Compose

Goals and Results

- The ability to analyze data at scale irrespective of details such as origin or format poses unique challenges but tremendous rewards.
- Additionally, a framework that simplifies the integration process to pointing at a data source and tying it to a rules engine provides large gains in design efficiency from an engineering perspective.
- Furthermore, if this system can be used in fields that are as domain-specific as bioinformatics and then also generalized, it would provide a significant proof-of-concept for linked data applications.
- SemNExT has already been applied in identifying an additional layer in the cortical development cycle, the Neuroectoderm (orange in the diagram above) [7]. It will continue to be used in exploring the relationships between genes as they relate to the expression of diseases and other disorders, and hopefully in other unique and varied domains.

Glossary:

- RPI - Rensselaer Polytechnic Institute
- ReDrugS - Repurposing Drugs with Semantics
- SemNExT - Semantic Numeric Exploration Technology
- UMLS - Unified Medical Language System
- JSON-LD - Javascript Object Notation for Linking Data
- SPARQL - SPARQL Protocol and RDF Query Language
- TWC - Tetherless World Constellation
- UDD - University Data Directory
- UVD - University Virtual Directory

Acknowledgements:

We would like to recognize the following individuals for their contributions and the immeasurable value they’ve added to this project.


Affiliations:

- Rensselaer Polytechnic Institute, Troy, NY 12180, USA
- Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
- Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
- Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
- National Science Foundation, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
- Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

References: