Digital Enterprise Research Institute

Cataloguing and Linking Life Sciences LOD Cloud

Ali Hasnain, Stefan Decker, and Helena Deus

{ali.hasnain, stefan.decker,helena.deus}@deri.org

Abstract

The Life Sciences Linked Open Data (LSLOD) Cloud is currently comprised of multiple datasets that add high value to biomedical research. However, navigating these multiple datasets is not easy as most of them are fragmented across multiple SPARQL endpoints, each containing trillions of triples and represented with insufficient vocabulary reuse. To retrieve and match from multiple endpoints, the data require to answer meaningful biological questions, it is first necessary to catalogue the data represented in each endpoint. We explore the schema used to represent data from a total of 52 meaningful Life Sciences SPARQL endpoints and present our methodology for linking related concepts and properties from the "pool" of available elements. We found the outcome of this exploratory work not only to be helpful in identifying redundancy and gaps in the data, but also for enabling the assembly of complex federated queries. We present three different approaches used to weave concepts.

Introduction

To achieve the ability for assembling queries encompassing multiple graphs hosted at various places, it is necessary either that vocabularies and ontologies are reused or that translation maps between the different terminologies are created. Two approaches for enabling integrated queries over LSLOD:

1. "***a priori integration***", relies on linked data representations schemas that make use of the same vocabularies and ontologies.

2. "***a posteriori integration***", facilitate mapping rules between different schemas, enabling the modification of the topology of queried graphs and the integration of datasets even when alternative vocabularies are used.

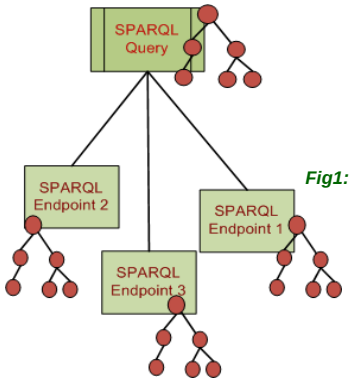

Fig1: "a-priori integration"

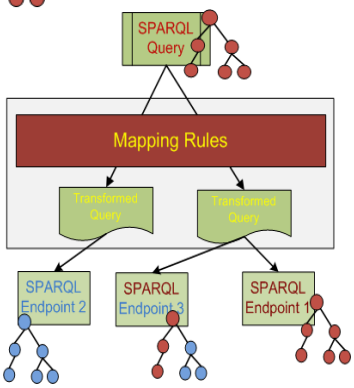

Fig 2: "a posteriori integration"

Method

We catalogued the LSLOD by harvesting, from 52 SPARQL endpoints the set of distinct concept/properties that may be used to query the data and the resulting triples were organized in an RDF document, the LSLOD Catalogue.

The LSLOD catalogue resulted in a "pool" of 12,396 concepts and 1,255 distinct properties from 52 endpoints. We combined several approaches for creating links between concepts and properties and resulted into 3 types of matching:

1. ***Naïve_Matching***:
   `type(I2,D1):=type(I1,D1),type(I2,D2),label(D1,L),label(D2,L)`
   -where I1 and I2 are instances; D1, D2 are two concepts, and L is the shared label.

2. ***Named Entity Matching:*** Similar pattern matching.

3. ***Domain Matching:*** `map(D1 , D2) := type(I1, D1), type( I2, D2), hasKey(D1, inchi1), hasKey(D2, inchi2), same (inchi1, inchi2)`

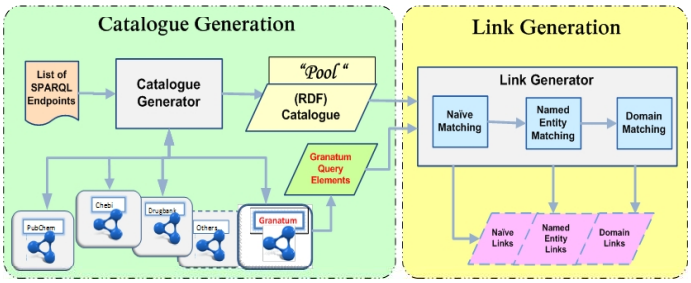

Fig. 3. Architecture of the components involved in LOD Catalogue and Link development

Results

In our initial exploration of LSLOD we found a total of 12,396 concepts, of which 12,119 were unique and out of 40,833 properties, 1,255 of which were unique

Total Concepts	12396	% Distinct
Total Distinct Concepts	12119	2.2%Reused
Semi-auto NamedEntity	11343	93.6 %
Domain Match	248	2.0 %
Naïve Match	92	0.8 %
Unmapped	402	3.5 %

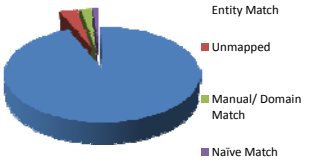

Fig 4: Matched Concepts Statistics

Total Properties	40833	
Distinct Properties	1255	96.9%reused
Naïve Linking	149	11.8%
Manual Linking	400	31.8%
Out-of-scope/Unlinked	706	56.2%

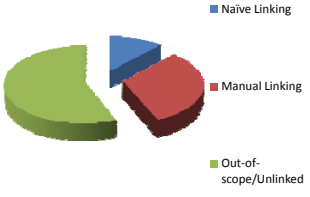

Fig 5: Matched Properties Statistics

Enabling **Networked** Knowledge

Acknowledgement This work has been funded by an EU- project "*Granatum*" which is ppartially funded by the European Commission under the Seventh Framework Programme in the area of Virtual Physiological Human (ICT-2009.5.3).