Cataloguing and Linking Life Sciences LOD Cloud

Ali Hasnain, Ronan Fox, Stefan Decker and Helena F. Deus

Digital Enterprise Research Institute (DERI) NUI Galway, Ireland
{ali.hasnain,ronan.fox,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. The ability to navigate through these datasets in order to derive and discover new meaningful biological correlations is considered one of the most significant resources for supporting clinical decision making. 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 required to answer meaningful biological questions, it is first necessary to catalogue the data represented in each endpoint, in order to understand how powerful queries traversing several SPARQL endpoints can be assembled. In this report, 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. In this report we present three different approaches used to weave concepts and properties and discuss their applicability for creating complex links in the LSLOD cloud.

Keywords: Linked Open Data, SPARQL, Life Sciences, Query Element.

1 Introduction

In the past few years, the linked open data cloud has earned a significant attention and it is becoming the *de facto* standard for publishing data on the web[1]. One of the ambitions behind the linked data effort is the ability to create a web of interlinked data which can be queried using a unified query language and protocol, regardless of where the data is stored. The life sciences domain has been one of the early adopters of linked data, and a significant portion of the linked data cloud is comprised of datasets from this domain, including multiple datasets from the *bio2rdf*[1] project, *linkedlifedata*[2], the *health care* and *life sciences knowledge base* [3](HCLS Kb), *neurocom-*

1http://bio2rdf.org/

2http://linkedlifedata.com/

3http://www.w3.org/TR/hcls-kb/

adfa, p. 1, 2011.

mons[4]and the *linked open drug data effort*[5]. These efforts have been partially devised and are still motivated by the deluge of data in biomedical facilities in the past few years, which is still in need of a single programmatic interface to access and query any life sciences dataset regardless of its representation formalisms. Although the publication of datasets as Linked Data is a necessary step towards achieving unified querying of biological datasets, it is not enough to achieve the interoperability necessary to enable a queryable web of life sciences data since it solves only the syntactic interoperability problem without addressing the interoperability problem caused by use of multiple overlapping terminologies when representing the data[2],[3]. 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[4]. As discussed in[5], there are two approaches that can be considered for enabling integrated queries using Linked Data: "*a priori integration*", which relies on linked data representations schemas that make use of the same vocabularies and ontologies and "*a posteriori integration*", a methodology that makes use of 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. As an example, there are multiple datasets in the LSLOD describing the concept "Molecule"- in *Bio2RDF'skegg*dataset, they are represented as chemical compounds using <*kegg#Compound*> whereas in *chebi*, these are identified as <*chebi#Compound*>and in *BioPax* they are denoted as <*biopax-level3.owl#SmallMolecule*>. A "*posteriori integration"* enables retrieving instances from these three concepts using a single triple pattern, provided they are mapped to each other and the SPARQL engine enables query transformation (e.g. SWobjects[6]). Using *a posteriori integration*, the following SPARQL algebra would enable the query rewrite necessary to retrieve all instances of "Molecule":

```
CONSTRUCT (bgp(triple?molecule
agr:Molecule))unionService(<kegg/sparql>,<kegg/sparql>bgp(triple ?mole-
cule a<kegg#Compound>)))
Service(<chebi/sparql>,<chebi/sparql>
bgp(triple ?msolecule rdf:type <chebi#Compound>))))
```

The "*a posteriori*" approach thereforerelies on identifying and creating the rules to transform the topology of the graphs using "*CONSTRUCT*" templates, thus enabling integration even when terminologies are not reused. *A posteriori* solutions are favored by Semantic Web Technologies given the extensive standardization of mechanisms to support the assignment of instances to new concepts through inference by simply creating a link that describes, for example, that two concepts are "*subClassOf*" of each other [5].Identifying and creating such links between similar or related concepts and properties is therefore a key requirement for the *posteriori* approach. In this report, we will focus on linking approaches that enable "a *posteriori integration*" in the LSLOD.

4http://neurocommons.org/page/Main_Page

5http://www.w3.org/wiki/HCLSIG/LODD

Currently a few *Query Engines QE* exist to support several techniques developed to meet the requirements of efficient query computation in the distributed environment. FedX, for example[7], is a project which extends the *Sesame Framework* with a federation layer that enables efficient query processing on distributed Linked Open Data sources. *SWobjects*[8] is a query engine and "*swiss-army-knife*" of semantic web technologies which supports also the query transformation requirement for *"a post-eriori"* data integration.

1.1 Challenges in linking LSLOD

To enable a posteriori integration of LSLOD, we introduce an approach to discover related and/or similar linked life sciences concepts and properties to facilitate federated SPARQL queries. We do not attempt to create a new semantic matching algorithm but to apply existing ones. In addition to the "*Compound*" example mentioned earlier (previous section), there are other concepts which instances should be returned upon querying "*Molecules*" such as instance of" *Drugs"* or "*Ingredients"* (Figure[6]), which can be considered as *subClassOf* "*Molecule*". This type of problem is prevalent in LSLOD. Furthermore, it is conventional in LSLOD for multiple endpoints to contain different fractions of data or predicates about the same entities – as an example, the *chebi* dataset would contain information related to the mass or the charge of a *Molecule*, whereas the *kegg* dataset contains information about the *Molecule*'s interaction with biological entities such as *Proteins*. There is an utmost need for linking particular concepts to address the issue of data inconsistency and heterogeneity. In this report we present our approach for linking concepts/classes and properties available at different SPARQL endpoints. Our approach will be discussed in detail in section 3.

2 Related Work

The studies or categories that are related to the work presented here come under the topic of "linking heterogeneous data". In this report we considered all the SPARQL endpoints in LSLOD to be represented according to a schema (list of concepts and properties).One typical way of addressing the data heterogeneity problem is through usage and alignment of ontologies. Semantic information systems use ontologies to represent domain-specific knowledge and support its users by enabling the usage of ontology terms to represent data and construct queries[9]. A system named B LOOMS was presented in [10] for finding schema-level links between LOD datasets. In BLOOMS, ontology alignment was achieved by bootstrapping information already available in the LOD cloud. BLOOMS relies heavily on Wikipedia and also on the API used for the ontology alignment [11]. Although BLOOMS would potentially be a candidate for achieving the linking proposed here, the life science domain is not covered in sufficient detail in Wikipedia's categorization to render sufficient and useful

[6] http://hcls.deri.org/RoadMapEvaluation/#sameData

links. Furthermore, ontology alignment approaches cannot be applied in our case given that such methods frequently rely on a set of assumptions that our datasets did not follow: 1) they often require an ontology as a starting point whereas the majority of SPARQL endpoints explored here did not have a clear representation ontology; 2) they do not attempt to link beyond the label or concept name, whereas named entity matching can provide a powerful method for linking and 3) they do not take into consideration domain matching approaches which are the core of our research and will be discussed in detail in the next sections. In [23] an alternative data linking mechanism is presented that relies on the unsupervised discovery of the required similarity parameters while taking into account several desired properties instead of relying only on labeled data. Ontology alignment techniques are more suited towards solving integration challenges where data has been structured as a hierarchy, which is not the case in the majority of SPARQL endpoints considered. The Silk Framework [12]provides a language for specifying and creating links between entities by matching based on predicates. Although we made use of it on the course of this work, it is not automated enough to cover the needs described as they still rely on extensive configuration and participation of data producers and consumers with expert knowledge. Also, it does not provide a mechanism for linking schema elements. The Silk framework can be useful for creating syntactic links (naïve) as it uses string comparison algorithms. The *VoID* vocabulary[13],which we used in this work is helpful for enabling the description of Linked Datasets, but cannot provide an automated way for link discovery between LOD datasets. The SameAs.org service also addresses the problem partially at concepts level by finding co-references between different data sets.

3 Methodology

We have catalogued the LSLOD by harvesting, from SPARQL endpoints and the set of distinct concept/properties that may be used to query the data. A total of 52 different SPARQL endpoints[7]were catalogued and the resulting triples were organized in an RDF document, the LSLOD Catalogue. These 52 endpoints include publically available bio2rdf datasets and datasets in *CKAN*[8] tagged with "*life sciences*" or "*healthcare*". Concept and properties were obtained by issuing queries such as *"select distinct*where{[]a?concept}*"and "*select distinct* where{<URI>?property ?object}"*.

As explained in the previous section, the LSLOD contains multiple overlapping instances which are described using different terminologies (e.g. *bio2rdf:compound*vs*biopax:smallMolecule*). However, for each data consumer, there is typically a set of preferred keywords (e.g. *Molecule, Name, Weight*) for retrieving these instances from LSLOD; we define these keywords as query elements (*Qe*)[9].In this report we present our 3-prongued approach for creating schema-level links between concepts and properties in the LSLOD catalogue and a predefined list of *Qe*. For queries expressed using schema keywords, this schema-level linking is meant to

7http://hcls.deri.org/RoadMapEvaluation/#Sparql_Endpoints

8http://wiki.ckan.org/Main_Page

9http://hcls.deri.org/RoadMapEvaluation/#Query_Elements

support the translation of SPARQL queries, assembled using a set of user-defined *Qe*, into a syntax that is amenable for querying multiple endpoints. For example when the data consumer requests instances of type "*Molecule*", the topology of remote graphs containing such instances should be transformed, according to a set of rules, into a compatible syntax in the remote service (e.g. instances of "*Compound*" is the equivalent query in the *CHEBI* endpoint). Our aim was to enable the automated identification of such rules and represent them in a format amenable for use by a federated SPARQL engine.

In the following sections, we describe weaving the "concepts" and "properties" in the LSLOD catalogue to a list of *Qe* determined as relevant by a set of life sciences experts on the course of the *EU GRANATUM* [10]project, focused in the domain of cancer chemoprevention. It is worth noticing that this initial set of *Qe* is a subset of the LSLOD catalogue, collected from the *GRANATUM SPARQL* endpoint[11], resulted in *Qe* e.g. *gr:Molecule, gr;Protein* and thus it can be replaced with any subset of the LSLOD catalogue. In order to enable the query rewrite necessary for "*a posteriori*" integration queries, links between these *Qe* and concepts/properties in LSLOD were created using predicates that can be used by reasoning engines to infer new instances/links (e.g. *subClassOf* and *subPropertyOf*). Doing so ensures that a federated query engine is able to make use of RDFS reasoning to transform a simple query into a federated query.

3.1 Links Creation

The LSLOD catalogue resulted in a "pool" of 12,396 distinct concepts and 1,255 distinct properties from 52 endpoints, the majority of which were not included in any Bioportal[12] ontologies. Also, as described in the related work section, none of the existing semantic matching/ontology alignment approaches could be used for addressing the challenge in its entirety (results in the next section). As such, we combined several approaches towards maximizing the number of links created. We divided our approach into 3 types of matching, described below namely: 1) *Naïve*; 2) *Named Entity*; 3) *Domain dependent*.

Naïve Matching/Syntactic Matching/Label Matching.

The simpler method of creating links between concepts is through naïve matching. In our catalogue development phase we captured the labels of all the concepts and properties or, when labels were missing, a new label was created from the last portion of the URIs. In the majority of cases, when two concepts share a label (e.g. "Compound"), they can confidently be linked together in same context(e.g. in LSLOD a "*Compound*" is always a "*Chemical Compound*"). The algebra used by a SPARQL federated engine to assign instances to the naïve matched concepts is formalized as:

`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.

10 http://www.granatum.org/

11 http://hcls.deri.org:8080/openrdf-sesame/repositories/granatumLBDS

12http://bioportal.bioontology.org/#

Named Entity Matching.

A significant number of instances in LSLOD are annotated to concepts representing the same entity (e.g. Molecule) but differ in their labels (e.g. Compound or Drug).Given that our main concern was to enable query transformation, this method was also used for matching concepts even when they are not exactly the same: as an example, even though "*Compound*" and "*Drug*" are not always synonyms, instances of "*Compound*" and "*Drug*" are also the instances of "*Molecule*". For such links, we created "*bags of related words*" through synonym and related terms identification. *WordNet* [14], and *Unified Medical Language System(UMLS)*[15]vocabularies were used to achieve automated similarity and relatedness scores with limited success (non specific, un-realistic and redundant links). To improve and explain this observation we contacted domain experts after examining the unmatched concepts; the involvement of domain experts was minimal and only used for filling gaps in recognizing identifiable patterns e.g.<http://bio2rdf.org/blastprodom:PD002610>and all concept URIs with similar patterns could be mapped to the *Protein Qe* automatically[13].

Manual and Domain dependent unique identifier Matching.

Domain matching relied on *properties that uniquely identify*[14] concepts. For example, *InChi*[15]is a property specifically devised for describing molecules. We captured these properties using *owl: hasKey*. In addition to enable schema-level matching, identifying these properties has the added advantage of enabling the automated linking of instances as well. The following formalization describes the assignment of instances to two concepts matched using domain matching:

```
map(D1 , D2) := type(I1, D1), type( I2, D2), hasKey(D1, inc-
hi1),hasKey(D2, inchi2),  same (inchi1, inchi2)
```

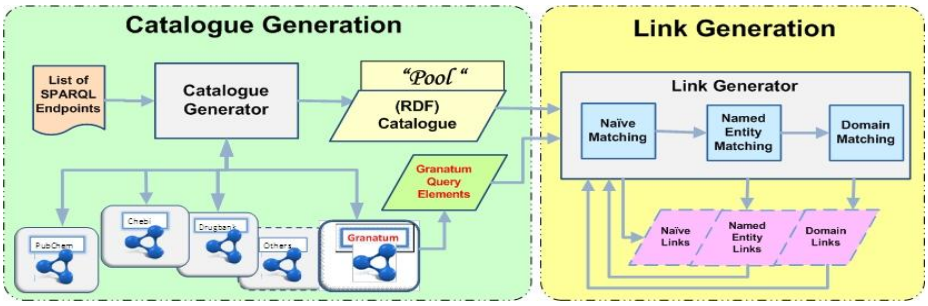

Figure. 1. Architecture of the components involved in LOD Catalogue and Link development

Figure 1 presents the complete overview of our method. A LSLOD catalogue is created given a list of SPARQL endpoints. Given a selection of *Qe* identified by the data consumer, concepts and properties from the "pool" were linked to the *Qe* by

13http://hcls.deri.org/RoadMapEvaluation/#Similar_Class_Patterns

14http://hcls.deri.org/RoadMapEvaluation/#Domain_Matching_Property

15http://www.iupac.org/home/publications/e-resources/inchi.html

applying all the approaches sequentially (figure 1). The linking process was designed to be iterative and to reuse the output of the previous stages, i.e. after each linking stage all the previous stages were repeated using the new set *Qe* matches resulting from previous stages. The RDF catalogue and the matched concepts can be examined at publicly available endpoint `(http://srvgal78.deri.ie/arc/roadmap.php)`

4 Results and Discussion

In this section we present our results and findings related to LSLOD catalogue development and Link Creation. 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. In section 4.1 we expose and discuss our linking results obtained with *WordNet.* Section 4.2 discusses the statistics regarding the mapping created on the basis of different sequential approaches discussed in section 3.1.

4.1 Linking results using *WordNet*

WordNet similarity measures can be classified into three categories: *Edge counting-based; Information content-based* and *feature based*. Using *WordNet* thesauri we attempted to automate the creation of *bags of related words* using 6 algorithms: *Jing & Conrath*[16], *Lin*[17], *Path*[18],*Resnik*[19],*Vector*[20]and *WuPalmer*[21].Concepts were considered the same when the similarity/relatedness value between them was *"1"* and dissimilar/ unrelated when *"0";* other threshold values were also considered. The links created were evaluated by the experts working in *Granatum*.

The linking results obtained by usage of these algorithms are available in figure2.The results show that a negligible amount of links was created with a maximum of 1.08% accepted links using Resnik. Only the links that have similarity greater than *0.9*were considered correct; links to more than one *Qe* via Resnik needed manual intervention to decide which link is appropriate.

Algorithms	Total Link	% link created	Correct Links	% correct links
Jing/Conrath	30	0.247545177	21	0.173281624
Lin	104	0.858156614	41	0.338311742
Path	18	0.148527106	18	0.148527106
Resnik	3592	29.63940919	131	1.080947273
Vector	18	0.148527106	18	0.148527106
WuPalmer	647	5.338724317	107	0.882911131

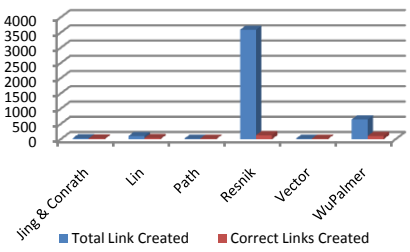

Figure. 2. *WordNet* based linked concepts statistics

These results suggest that the concepts and properties in LSLOD are too specific and thus *WordNet* thesauri is possibly too generic for this domain. We had similar results with the *UMLS* thesaurus, where linking results were not considered relevant as very low similarity scores were assigned for reasonably similar concepts (full results table at[16]) - e.g. UMLS similarity(molecule, drug) = 0.4835 - with high similarity scores

16http://hcls.deri.org/RoadMapEvaluation/#UMLS

being assigned for some dissimilar concepts - e.g. UMLS similarity(*molecule, organism*) = 0.7298. Also, we found that several concepts which labels consisted of compound words such as underscore, camel-case or dash separated words (e.g. *PathwayDatabase*) could be easily used for matching using the simplest strategy (Naïve Matching). Most of the existing tools and technologies do not support this heterogeneity in label composition and this could also explain the poor matching results obtained. It is therefore *Wordnet* or *UMLS* couldn't be considered as a basis for our proposed linking approaches, results of which are presented in next section.

4.2 Concept linking results

The link creation results from the various matching approaches are available in Figure 3 – the large majority of concepts (93.6%) were mapped using the Named Entity Matching approach. One of the reasons for explaining these results, points to the differences in the methods used to populate the SPARQL endpoints. In the majority of SPARQL endpoints, the number of concepts retrieved was low (between 1 and 108), while in two cases (PDB and SGD), the number of concepts retrieved was significantly larger (1672 and 9476, respectively). We noticed that, in these cases, concepts, as opposed to instances, were being used to describe entities of type Molecule or Organism, with each concept containing only one or two instances. Our methodology could map these concepts based on named entity matching (e.g. concepts with pattern *http://bio2rdf.org/hmmpir*: could be mapped to *Protein*). However, in future work we will transform the topology of the graphs in each of these two endpoints in order to match graphs with the same topology.

In many cases the URI representing concepts consisted of url-encoded labels (e.g *http://bio2rdf.org/pdb:1%2C1%2C5%2C5tetrafluorophosphopentylphosphonicAcidAdenylateEster)*,which made linking a challenge. A similar situation was found when the concept was formed using alpha-numeric combination for which a label could not be found either in its source SPARQL endpoints or through browsing ontology registry services such as *Bioportal* , e.g. *<http://bio2rdf.org/so:0000436>*.

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

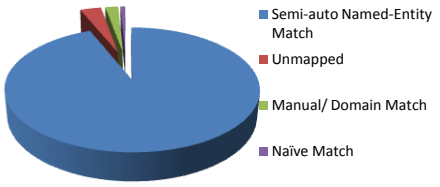

Figure. 3. Linked concepts statistics

Domain experts were only consulted for filling the gap in recognizing possible matching patterns and the subsequent processes were automated. Although a very low percentage of linking was achieved through the naïve matching or domain matching, the quality of these links was very high as it relied on the precise, often standardized terms, when available. There were cases when a concept could not be directly mapped to any of our query elements but nevertheless remained within scope of the query elements. As an example, we found instances of concept "*Peptide*", which represent a portion of a "*Protein*" but not necessarily an instance of a "*Protein*". In those cases,

the appropriate relationship would to specify that *{ Peptide containedIn: gr:Protein }* or to lift the term "*Peptide*" to query element level. Automating the creation of such links was out of scope for this report.

4.3 Properties linking results

Out of 40833 properties, 1255 were distinct, indicating that there is significantly more property reuse in LSLOD than concept reuse. Figure4 illustrates the links created by each of the approaches described. Approximately 56.2% of the properties remained unlinked due to the fact that a significant proportion of properties were considered irrelevant in the context of the *Qe* selected, which indicates that they may be applicable with a new set of *Qe*.

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

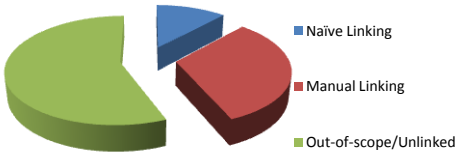

Figure. 4. Linked properties statistics

It is also worth noticing that 3.5% of identified concepts and 56.2% of the properties remained unlinked; they were either out of scope or could not match any of the query element. This indicates that quality, as well as the quantity of links created is highly depended on the input list of query elements. The larger the number of query elements, the better the quality of the links created e.g a concept called *?peptide* would link to *gr:Peptide* if such element existed. However, selecting from a large number of query elements is not always practical for querying. As such, there should be a balance between the number of *Qe* and the accuracy of links created. We attempted to link almost 12000 concepts to 60 query elements to avoid the need to browse more than 12000 concepts in order to assemble a query. Notice that the query elements selected were chosen by domain experts; a change in the query element set may result in different links being created and therefore our solution can be applicable for any set of query elements.

5 Conclusion and Future Work

Our preliminary analysis of existing LSLOD SPARQL endpoint reveals that the schemas of most datasets cannot be easily linked together using existing approaches. In fact, in the majority of cases there is very little ontology and URI reuse. In this report we describe a combined approach for data linking to facilitate "*a posteriori*" SPARQL query transformation in the LSLOD. Our methodology relies on systematically issuing queries on various life sciences SPARQL endpoints and collecting its results in an approach that would otherwise have to be encoded manually by domain experts or those interested in making use of the web of data towards answering meaningful scientific questions. Our aim was to support "*a posteriori integration*", i.e. integration of instances in LSLOD where different terminologies were used. As a result of this work, rules such as R1:*{ chebi:Compound rdfs:subClassOf gr:Molecule*

} can be used by a query engine supporting federated query to transform the query *[?molecule a gr:Molecule]* into the federated alternative *[{?molecule a* chebi:Compound } UNION {*?molecule* a *gr:Molecule}]* since the integration of R1 ensures that all instances of *chebi:Compound* are also instances of *gr:Molecule*whereas the opposite is not true [14].

Created Links were evaluated by the experts working in Granatum and in future, will be evaluated by experts working in this area to evaluate whether the results of the queries returned are actually what the scientist would be looking for. A possible extension to the linking work presented here is the implementation of corpus based similarity using *Wikipedia-based Explicit Semantic Analysis* ESA Measure [22].Our work has been limited to the linking of concepts and properties based on the terminological matching using three pronged approaches, hence will be extended in future beyond terminological linking approaches.

References.

1. C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009.
2. J. Quackenbush, "Standardizing the standards," *Molecular systems biology*, vol. 2, no. 1, 2006.
3. S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, and C. Goble, "Why linked data is not enough for scientists," *Future Generation Computer Systems*, Aug. 2011.
4. A. Polleres, "Semantic Web Technologies : From Theory to Standards."
5. H. F. Deus, E. Prud'hommeaux, M. Miller, J. Zhao, J. Malone, T. Adamusiak, J. McCusker, S. Das, P. R. Serra, R. Fox, and M. Scott Marshall, "Translating standards into practice - One Semantic Web API for Gene Expression.," *Journal of biomedical informatics*, Mar. 2012.
6. E. Prud'hommeaux, H. Deus, and M.S.Marshall, "Tutorial: Query Federation with SWObjects," 2011.
7. A. Schwarte, P.Haase, K.Hose, R.Schenkel, and M.Schmidt, "FedX: a federation layer for distributed query processing on linked open data,"*The Semanic Web: Research and Applications*,pp.481–486, 2011.
8. E. Prud'hommeaux, "SWObjects." [Online]. Available: http://www.w3.org/2010/Talks/0218-SWObjects-egp/#(1)[Accessed: 03 June 2012] .
9. M. Petrovic, I. Burcea, and H. A. Jacobsen,"S-topss: Semantic toronto publish/subscribe system,"in *Proceedings of the 29th international conference on Very large data bases-Volume 29*, 2003.
10. P. Jain, P. Hitzler, A. Sheth, K. Verma, and P. Yeh, "Ontology alignment for linked open data," *The Semantic Web--ISWC 2010*, pp. 402–417, 2010.
11. J. Euzenat, "An API for ontology alignment," *The Semantic Web--ISWC 2004*, pp. 698–712, 2004.
12. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Discovering and maintaining links on the web of data," *The Semantic Web-ISWC 2009*, pp. 650–665, 2009.
13. R. Cyganiak and M. Hausenblas, "Describing Linked Datasets-On the Design and Usage of voiD, the'Vocabulary of Interlinked Datasets'," 2009.
14. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An online lexical database*," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.
15. O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.
16. J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *Arxiv preprint cmp-lg/9709008*, 1997.
17. D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th international conference on Machine Learning*, 1998, vol. 1, pp. 296–304.
18. R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 19, no. 1, pp. 17–30, 1989.
19. P. Resnik, "Disambiguating noun groupings with respect to WordNet senses," in *Proceedings of the Third Workshop on Very Large Corpora*, 1995, pp. 54–68.

20. S. Patwardhan and T. Pedersen, “Using WordNet-based context vectors to estimate the semantic relatedness of concepts,” in *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*, 2006, vol. 1501, pp. 1–8.
21. Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994, pp. 133–138.
22. E. Gabrilovich and S. Markovitch, “Computing semantic relatedness using wikipedia-based explicit semantic analysis,” in *Proceedings of the 20th international joint conference on artificial intelligence*, 2007, vol. 6, p. 12.
23. *Nikolov, Andriy; d'Aquin, Mathieu and Motta, Enrico. Unsupervised learning of link discovery configuration In:9th Extended Semantic Web Conference(ESWC 2012),27-31 May 2012,Heraklion, Greece*