

Expanding the Pathway and Interaction Knowledge in Linked Life Data

Vassil Momtchev¹, Deyan Peychev¹, Todor Primov¹, Georgi Georgiev¹,

¹ Ontotext AD, Tsarigrasko Shosse. 135,
1784 Sofia, Bulgaria
{first.lastname}@ontotext.com

Abstract. Linked Data already gained popularity as a platform for data integration and analysis in the life science and health care domain. This paper is an ongoing report for the recent developments in the Linked Life Data platform and the Pathway and Interaction Knowledge Base (PIKB) dataset. They integrate semantically molecular information and realize its linking to the public data cloud. The dataset interconnects more than 20 complete data sources and helps to understand the "bigger picture" of a research problem by linking unrelated data from heterogeneous knowledge domains. To make efficient usage of the public linked data cloud, we have created instance alignment patterns that restore missing information relationships. As a final step, a massive number of semantic annotations (optimized for high recall or precision) is generated between the linked data instances and the unstructured information. The LDD prototype is available at <http://linkedlifedata.com>.

Keywords: linked data, data integration, pathways, life sciences, RDF

1 Introduction

In recent years we have witnessed a huge explosion of biological, medical, and chemical data in terms of volumes and heterogeneity. Data integration continues to be a serious bottleneck for the expectations of increased productivity in the pharmaceutical and biotechnology domain.

In a single enterprise, the researchers require different views over one and the same data. During the entire process, the research team has to: identify potential molecular species (genes and proteins of interest) and screen their molecular properties; analyze the molecular interactions in the context of cellular and physiological processes; mine huge amounts of structured and more often, unstructured, textual information for pharmacological and clinical data. The analysis of molecular interaction data is usually limited to the interpretation of the different types of interaction and biological pathways in the context of cellular and physiological processes. These limitations are determined by the data integration methods used for the generation of the interaction knowledge bases. But to understand the "bigger picture" of a research problem, the scientists often need to link visually unrelated data from heterogeneous knowledge domains.

Semantic Web technology seems to be a promising technology for reducing the complexity of combining data from multiple sources and resolving classical integration problems related to the information accessibility. In the literature, there are several examples that apply the RDF technology as "semantic glue". [3] summarizes the different approaches as centralized (Bio2RDF [2], the HCLS Knowledge Base [5]) and distributed. Despite the significant advantages of the presented approach in [3], we believe that it will not be possible the efficient execution of complex real-life queries, which requires merging of remote datasets.

Linked Life Data is a data integration platform that realizes a massive RDF warehouse solution extended with inference and semantic annotations support. Its back-end is the OWLIM semantic repository [10] that is proven to scale up to 20 billion RDF statements [7].

2 Input Datasets

Since the beginning of the Semantic Web, many bioinformatics and biomedical resources announced RDF versions of their distribution or used the technology for semantic data integration. Recently, a number of public services such as Uniprot RDF, Bio2RDF [2], LODD [6], etc. announced their compliance with the Linked Data best-practices by exposing the data for access via the HTTP protocol and emphasizing the interconnections and relationships of this data. PIKB is a semantically integrated dataset that links to the public cloud pathway, interaction, gene, protein, bibliographic and biomedical thesauri knowledge. It generates connections to a number of sources like Uniprot and LODD.

We implemented the RDF representation of the PubMed, UMLS, Entrez-Gene, and OBO Foundry data sources. For PubMed and Entrez-Gene we strictly followed the database schema. The UMLS dataset is limited only to vocabularies with Category 1 and Category 2 license type (e.g., SNOMED and other high-quality resources are omitted because of their strict licensing policy) and a SKOS representation is generated using a custom script. All OBO ontologies are transformed to SKOS schema according the guidelines proposed by [9]. Table 1 presents all datasets included in PIKB that are available for downloading from the LLD website.

Table 1. Pathway and Interaction Knowledge Base data sources.

Data source	Statements (explicit)	Schema	Description
Entrez-Gene	107,193,308	Custom schema	Genes and annotation
BioGRID	1,892,897	BioPAX 2.0	General Repository for Interaction Datasets
NCI /NPIDb	333,415	BioPAX 2.0	Human pathway interaction database
The Cancer Cell Map	173,914	BioPAX 2.0	Cancer pathways database
Reactome	2,538,793	BioPAX 2.0	Human pathways and interactions
INOH	432,456	BioPAX 2.0	Pathway database
HPRD	1,805,651	BioPAX 2.0	Human Protein Reference Database
HumanCyc	341,225	BioPAX 2.0	Encyclopedia of Human Genes and Metabolism
IMID	154,408	BioPAX 2.0	Protein interactions extracted from the literature
IntAct	11,005,555	BioPAX 2.0	Protein interaction database
MINT	7,915,613	BioPAX 2.0	Molecular INTeraction database
KEGG	18,128,735	BioPAX 2.0	Molecular Interaction
PubMed ¹	807,851,455	Custom schema	Citations from Medline and other life sciences journals
UMLS semantic network	1,368	SKOS	Semantic categorization of terminology in multiple domains
UMLS meta-thesaurus	12,420,882	SKOS	Database that contains information about biomedical and health related concepts, their various names, and the relationships among them
Disease Ontology	446,066	SKOS	Controlled medical vocabulary designed to facilitate the mapping of diseases and associated conditions to particular medical codes such as ICD9CM, SNOMED and others.
Human Phenotype Ontology	70,911	SKOS	Human phenotype ontology
Symptom Ontology	4,163	SKOS	Symptoms ontology

Table 2 indicates other datasets loaded in LLD to maximize the value of the PIKB information.

¹ The dataset contains duplicated statements.

Table 2. Other data sources loaded in LLD.

Data source	Statements (explicit)	Schema	Description
Uniprot	1,146,084,021	Supplied by the provider	Protein sequences and annotations
DrugBank	493,794	Supplied by LODD	Chemical, pharmacological, and pharmaceutical drug data
SIDER	96,272	Supplied by LODD	Drug side affects
Diseasome	69,546	Supplied by LODD	Network of disorders and disease genes linked by known disorder–gene associations
Dailymed	116,992	Supplied by LODD	Information about marketed drugs
LinkedCT	7,035,974	Supplied by LODD	ClinicalTrials.gov represented into RDF
DBpedia ²	439,775,096	Supplied by the provider	Structured information from Wikipedia

3 LLD Design Decisions and Methodology

In this section we present important design decisions for the development process of the PIKB dataset and the integration methodology used in LLD.

3.1 URI Naming Conventions

Linked Data principles state that it must be possible to deference every URI and to access the related meta-data, [8]. Also we would like to keep the RDF format and its naming schema distributed by the original vendor to preserve the semantic interoperability with all tools and dataset that use them. However, the two ideas are in conflict with all generated RDF datasets that are not exposed via HTTP. In the case of the PIKB datasets, the only exception is the databases distributed in BioPAX format. We consider the cross tool interoperability more important and therefore stay with the following rule ordering:

- R1:** Preserve the original RDF structure if distributed by the owner.
- R2:** Use resolvable URIs for the data sources with no RDF distribution
- R3:** Construct the generated URIs in the form of `l1d:resource/db/type/id`
- R3:** Identify the graph names with `l1d:resource/db`
- R4:** Name all generated predicate URIs `l1d:resource/db/predicate`
- R5:** Generate stable new URIs based on unique label that describes the resource (see dataset provenance and updates)

² Modified version to remove cycles in the hierarchy.

3.2 Dataset Provenance and Updates

One of the major overheads in the warehouse systems is the information update. The RDF format is an abstract data representation model therefore the information synchronizations policy, when no incremental updates are available, follows a very straightforward procedure: 1) regenerate the data source 2) delete the existing graph information, and 3) import the new data. The simple update process, however, implies constraints over the way data is generated (see URI naming convention R5). For example, every resource identifier must be stable (e.g., it should not have been generated as a result of random function) in order to preserve all generated links to the resource. Another restriction is the need to separate the statement, created by independent processes (database loading, manual annotations, information extraction), into a separated graph

3.3 Linked Data Mapping

Extraction, transformation, loading (ETL) is a typical phase in the generation of every data warehouse. RDF warehousing requires similar operation to address the variety of different data modeling approaches. Based on more than 20 different RDF database representations, we have identified the following integration patterns to interconnect related resources. Figure 1 presents the Linked Data alignment process. The blue lines and the blue text of the captions (used either as part of the URI or literals) designate the criteria for linking the information. The specified mapping rules are not universally applicable for all RDF types and they are applied only to subsets of the information. The process of the subset selection and the rule application is manually controlled.

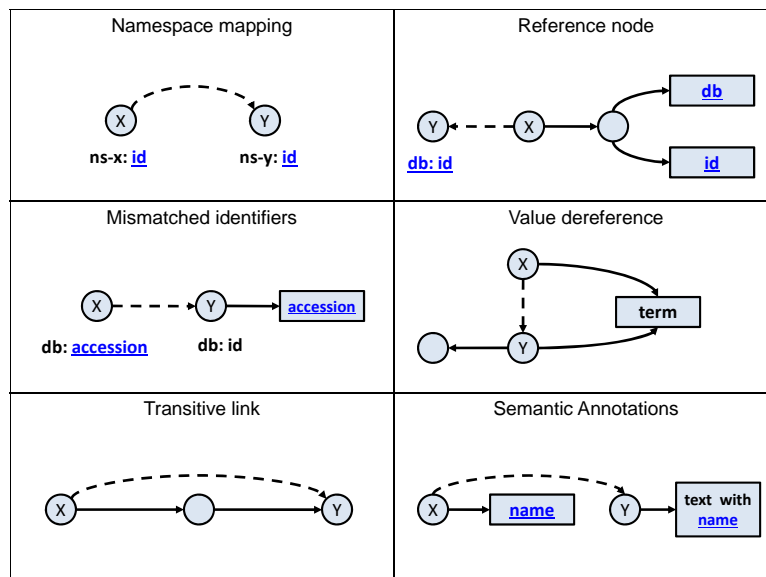


Figure 1. Linked Data instance alignment rules.

3.4 Semantic Annotations

In a nutshell, the term Semantic Annotation is used for assigning links between the entities, recognized by arbitrary information extraction algorithm, and their semantic descriptions. This sort of metadata provides class and/or instance information about the entities. Moreover, knowledge acquisition can be performed based on the extraction of more complex dependencies – analysis of relationships between entities, event and situation descriptions, etc. In essence, such metadata that is useful could be found in DBPedia [1], a promising community effort to transform Wikipedia. In the DBPedia data the predicate *wikilink* connects resources representing linked pages in Wikipedia. In LLD we use semantic annotation to add additional links between resources and to demonstrate the excellent multibillion statements scalability of the OWLIM repository. For example, in LLD the EntrezGene concept SP-A (entity describing surfactant associated protein A1) with URI `entrezgene:id/20387` contains GeneRIF with the following text:

SP-A is necessary for lungs to respond to hyperventilation or secretagogues with increased **DPPC** uptake and also modulates the PLA(2)-mediated degradation of internalized **DPPC**.

Our information extraction process recognizes the chunk DPPC (shown in bold above) as the concept 1,2-Dipalmitoylphosphatidylcholine, described in UMLS under the URI `lld:resource/umls/id/C0000039`. It should be noted that, although the name is 1,2-Dipalmitoylphosphatidylcholine, we resolve all aliases (alternative names) of this UMLS concept to the same resource. Finally, in our approach, the text is represented as a repository concept, e.g., the GeneRIF text shown above possesses the URI `entezgene:GeneRIF/208024`. This holds for any arbitrary text in our repository, e.g., PubMed article text, PubMed title, Entrez Gene description, etc.

4 Results and Discussions

The data in PIKB and the related datasets (shown in Table 2) is described by more than 2.217 billion explicit statements. The overall import took 40 hours on a standard server configuration. The average loading and inference speed varied between 5,000 and 60,000 statements per second, depending on the complexity of the loaded dataset (e.g., the size of the literals for full-text, inference complexity, and the degree of dataset interconnection). With no significant optimizations the LLD prototype and the underlying OWLIM engine seem capable of maintaining continuous updates and automating all post processing activities for datasets of a similar scale.

The Linked Data mapping rules are agnostic to the generated semantics. We have found no single efficient solution to implement them with a single component therefore their processing is divided as follows: 1) Namespace mappings in our particular dataset are all covered by the URI generation convention; 2) Reference node, Mismatched identifiers, and Value dereference are processed by a custom

implementation of Java reasoner; 3) Transitive links are declared as part of the OWLIM inference rules and OWL transitive properties; 4) Semantic annotations are processed by a GATE pipeline [4] executed in parallel. Table 3 presents statistics for the generated links between the data sources with the custom Java reasoner implementation.

Table 3. Linked Data mapping rules output

Source dataset	Destination dataset	Linked Data Mapping Rule	Number of connections	Semantic relationship
BioPAX (genes)	Entrez Gene	Reference Node	7,897	skos:closeMatch
BioPAX (GO)	GO (UMLS)	Reference Node	44,642	skos:relatedMatch
BioPAX (taxonomy)	NCBI Taxonomy (UMLS)	Reference Node	52,851	skos:closeMatch
BioPAX (proteins)	Uniprot	Reference Node	107,183	skos:exactMatch
Diseasome (gene)	Entrez Gene	Mismatched Identifiers	2,772	skos:closeMatch
Entrez Gene (GO)	Gene Ontology (UMLS)	Mismatched Identifiers	27,647	skos:exactMatch
Entrez Gene (taxonomy)	NCBI Taxonomy (UMLS)	Mismatched Identifiers	4,210	skos:exactMatch
Uniprot GO	GO (UMLS)	Mismatched Identifiers	48,652	skos:exactMatch
DrugBank (targets)	Entrez Gene	Value Dereference	1,515	skos:closeMatch

The named entity recognition process operates over a limited predefined list of literals and concept. Two different concept lists are tested. The first one (1,248,890 aliases) is optimized for a high precision result and the second (2,518,641 aliases) - for high recall. Consequently, two predicates are used: *mention* and *mentionStrict*. We selected more than 16 million literals containing concepts, where 15 million came from PubMed, as text for annotation. The named entity recognition process created 705,338,334 high recall semantic annotations, while the high precision approach generated 263,323,164.

The LLD platform demonstrates efficient search over highly heterogeneous and loosely coupled data. It is capable of executing queries that cover information from 7 different sources in a timely fashion. The platform and the PIKB dataset are used as a domain specific reporting tools for generating new information insights. The web front-end provides three paths to access the data: a web form for issuing SPARQL queries, a browser for exploring resources, and full text search in the graph containing the searched literal with matched resources.

As the main objective set for the system was to facilitate the mining of concealed relations among data, we have developed a step-by step scenario, which demonstrates the potential of the technology for interlinking information from multiple heterogeneous sources and for providing a more holistic view over a particular scientific problem.

We have prepared a set of predefined queries (they could be found and executed via the end user prototype interface, named “query 1” to “query 4”) that expand step by step the knowledge and, at the same time, increase the specificity of the asked question. The queries have the following pattern:

“Select all human genes (query 1), which code proteins with known molecular interactions (query 2) and are analyzed with molecular techniques (query 3). We can go even further in the mining as we can restrict the results just to gene/proteins, which are known drug targets (query 4)”

The last example (query 4) mines for information relations in 15 different data sources (Uniprot, Entrez-Gene, PubMed, DrugBank, BioGRID, NCI /NPID, The Cancer Cell Map, Reactome, INOH, HPRD, HumanCyc, IMID, IntAct, MINT, and KEGG) from 5 different biomedical domains (genes, proteins, molecular interactions, scientific publications, and drugs).

5 Acknowledgements

The work reported in this paper was partially supported by the EU FP7 IP 215535 LarKC. The authors wish to express special thanks to Bosse Anderson and the other researchers from AstraZeneca for their help with the use case and system specification. We wish to thank all implementers of the data sources (listed in alphabetical order) Bio2RDF, BioPAX, DBPedia, Entrez-Gene, LODD, Medline, Pathways Commons, and UMLS for making them publicly available.

References

1. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R. & Ives, Z.: DBpedia: A Nucleus for a Web of Open Data In: *The Semantic Web*, November (2008), S. 722-735.
2. Belleau, François; Nolin, Marc-Alexandre; Tourigny, Nicole; Rigault, Philippe & Morissette, Jean: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. In: *Journal of Biomedical Informatics*, Vol. 41, Nr. 5 (2008), S. 706-716.
3. Ben Vandervalk; Luke McCarthy; Mark Wilkinson, CardioSHARE: Web Services for the Semantic Web, *Semantic Web Challenge 2008*, 2008
4. Cunningham, H.: GATE, a General Architecture for Text Engineering In: *Computers and the Humanities*, Vol. 36 (2002), S. 223-254.
5. Health Care and Life Sciences Interest Group: A Prototype Knowledge Base for the Life Sciences, <http://www.w3.org/TR/hcls-kb/>
6. <http://esw.w3.org/topic/HCLSIG/LODD>
7. <http://www.ontotext.com/owlim/index.html>
8. <http://www4.wiwi.fu-berlin.de/bizer/pub/LinkedDataTutorial/>
9. Jupp, Simon. Bechhofer, Sean. Kostkova, Patty. Stevens, Robert. Yesilada, Yeliz. Document Navigation: Ontologies or Knowledge Organisation Systems? In *Network Tools and Applications in Biology (NETTAB'2007) - A Semantic Web for Bioinformatics: Goals, Tools, Systems, Applications*, June 2007
10. Kiryakov, Atanas; Ognyanov, Damyan & Manov, Dimitar: OWLIM - A Pragmatic Semantic Repository for OWL., Vol. 3807 Springer (2005), S. 182-192.