LarKC

The Large Knowledge Collider:
a platform for large scale integrated reasoning and Web-search

FP7 – 215535

D7a.2.1 Pathway and Interaction Knowledge Base

Coordinator: Vassil Momtchev
With contributions from: Deyan Peychev, Todor Primov, Georgi Georgiev ONTO

Document Identifier:	LarKC/2008/D7a.2.1 /v1.0
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	1.0
Date:	30.09.2009
State:	Final
Distribution:	Public

EXECUTIVE SUMMARY

Pathway and Interaction Knowledge Base is an ontology developed in the context of WP7a “Semantic Integration for Early Clinical Development”. The created ontology is a public resource that semantically integrates various free bioinformatics and biomedical data sources. For its evaluation and impact analysis we have chosen a community review process that will disseminate the knowledge base to a broader audience interested in the Semantic Web, Linked Data, life sciences and health care technologies. A description of PIKB ontology, the applied generation methodology and the implemented Linked Life Data (LDD) platform that hosts the knowledge are presented and submitted as paper to “Semantic Web Challenge 2009” [1] part of International Semantic Web Conference 2009 – Open Track [2].

DOCUMENT INFORMATION

IST Project Number	FP7 - 215535	**Acronym**	LarKC
Full Title	The Large Knowledge Collider: a platform for large scale integrated reasoning and Web-search		
Project URL	http://www.larkc.eu/		
Document URL			
EU Project Officer	Stefano Bertolo		

Deliverable	**Number**	D7a.2.1	**Title**	PIKB
Work Package	**Number**	WP7a	**Title**	Semantic Integration for Early Clinical Development

Date of Delivery	**Contractual**	M18	**Actual**	M18
Status	version 1.0		final □	
Nature	prototype □ report □ dissemination □ ontology X			
Dissemination level	public X consortium □			

Authors (Partner)	Vassil Momtchev, Deyan Peychev, Todor Primov, Georgi Georgiev, Rostislav Hristov (Ontotext)			
Responsible Author	**Name**	Vassil Momtchev	**E-mail**	vassil.momtchev@ontotext.com
	Partner	ONTO	**Phone**	

Abstract (for dissemination)	This report is supplementary to the M18 ontology deliverable Pathway and Interaction Knowledge Base (PIKB). The created ontology is a public resource that semantically integrates various free bioinformatics and biomedical data sources. For its evaluation and impact analysis we have chosen a community review process that will disseminate the knowledge base to a broader audience interested in the Semantic Web, Linked Data, life sciences and health care technologies. A description of PIKB ontology, the applied generation methodology and the implemented Linked Life Data (LDD) platform that hosts the knowledge is submitted as paper to “Semantic Web Challenge 2009” part of International Semantic Web Conference 2009 – Open Track.
Keywords	Linked Data, RDF warehouse, data integration, life sciences, early clinical development

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**

PROJECT CONSORTIUM INFORMATION

Participant's name	Partner	Contact
Semantic Technology Institute Innsbruck, Universitaet Innsbruck		Prof. Dr. Dieter Fensel, Semantic Technology Institute (STI), Universitaet Innsbruck, Innsbruck, Austria, E-mail: dieter.fensel@sti-innsbruck.at
AstraZeneca AB		Bosse Andersson AstraZeneca Lund, Sweden Email: bo.h.andersson@astrazeneca.com
CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA		Emanuele Della Valle, CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA, Milano, Italy, Email: emanuele.dellavalle@cefriel.it
CYCORP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O.		Michael Witbrock, CYCORP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O., Ljubljana, Slovenia, Email: witbrock@cyc.com
Höchstleistungsrechenzentrum, Universitaet Stuttgart		Georgina Gallizo, Höchstleistungsrechenzentrum, Universitaet Stuttgart, Stuttgart, Germany, Email: gallizo@hlrs.de
MAX-PLANCK GESELLSCHAFT ZUR FOERDERUNG DER WISSENSCHAFTEN E.V.		Dr. Lael Schooler Max-Planck-Institut für Bildungsforschung Berlin, Germany Email: schooler@mpib-berlin.mpg.de
Ontotext AD		Atanas Kiryakov, Ontotext Lab, Sofia, Bulgaria Email: naso@ontotext.com
SALTLUX INC.		Kono Kim, SALTLUX INC, Seoul, Korea, Email: kono@saltlux.com
SIEMENS AKTIENGESELLSCHAFT		Dr. Volker Tresp, SIEMENS AKTIENGESELLSCHAFT, Muenchen, Germany, E-mail: volker.tresp@siemens.com
THE UNIVERSITY OF SHEFFIELD		Prof. Dr. Hamish Cunningham, THE UNIVERSITY OF SHEFFIELD Sheffield, UK, Email: h.cunningham@dcs.shef.ac.uk

VRIJE UNIVERSITEIT AMSTERDAM		Prof. Dr. Frank van Harmelen, VRIJE UNIVERSITEIT AMSTERDAM, Amsterdam, Netherlands, Email: Frank.van.Harmelen@cs.vu.nl
THE INTERNATIONAL WIC INSTITUTE, BEIJING UNIVERSITY OF TECHNOLOGY		Prof. Dr. Ning Zhong, THE INTERNATIONAL WIC INSTITUTE, Mabeshi, Japan, Email: zhong@maebashi-it.ac.jp
INTERNATIONAL AGENCY FOR RESEARCH ON CANCER		Dr. Paul Brennan, INTERNATIONAL AGENCY FOR RESEARCH ON CANCER, Lyon, France, Email: brennan@iarc.fr
INFORMATION RETRIEVAL FACILITY		Dr. John Tait INFORMATION RETRIEVAL FACILITY Vienna, Austria Email : john.tait@ir-facility.org

TABLE OF CONTENTS

List of Acronyms

Acronym	**Description**
KB	Knowledge Base
LarKC	Large Knowledge Collider
LLD	Linked Life Data
PIKB	Pathway and Interaction Knowledge Base

1. Introduction

This report is supplementary to the M18 ontology deliverable Pathway and Interaction Knowledge Base (PIKB). PIKB and the access to it via Linked Life Data semantic data integration platform are public resources offered as a free service. Therefore the deliverable evaluation process is delegated to the community for a review. The paper is submitted to “Semantic Web Challenge 2009” – Open Track [1] collocated with the International Semantic Web Conference 2009 [2]. The paper describes the used input data sources, applied integration methodology, post-processing linked data instance alignments, and executed information extraction algorithms to generate the knowledge base. A pre-final copy is attached as part of the current deliverable report in appendix A.

2. References

[1] Semantic Web Challenge 2009 – http://challenge.semanticweb.org/

[2] 8th International Semantic Web Conference (ISWC 2009) – http://iswc2009.semanticweb.org/

Appendix A

Expanding the Pathway and Interaction Knowledge in Linked Life Data

Vassil Momtchev, Deyan Peychev, Todor Primov, Georgi Georgiev,

Ontotext AD, Tsarigrasko Shosse. 135,
1784 Sofia, Bulagaria
{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 Linked Life Data platform and the Pathway and Interaction Knowledge Base (PIKB) dataset that semantically integrate molecular information and realize its linking to the public data cloud. The dataset interconnects more than 20 complete data sources that help understanding the "big 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) are generated between the linked data instances and the unstructured information.

Keywords: Linked data, data integration, life sciences, health care, pathways, RDF, semantic annotations.

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 plans of increased productivity in the pharmaceutical and biotechnology domain.

The initial scientific research in the drug development process requires data mining in a complex array of knowledge domains - chemistry, biochemistry, microbiology, pharmacology, and medicine. In a single enterprise, the researchers require different views over one and the same data. For example, a scientist in the Drug Discovery phase will be much more interested in the molecular data, while his colleagues involved in the Pre-clinical Trials might be more interested in the pharmacological, physiological, and clinical information. 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. This more or less narrows the data model to be suitable just for a limited set of tasks. But to understand the "big 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 a "semantic glue". [2] summarizes the different approaches as centralized (Bio2RDF [1], the HCLS Knowledge Base [3]) and distributed. Despite the significant advantages of the presented approach in [2], we believe that it will not be possible to efficiently execute complex real-life queries, which requires merging of remote datasets (e.g., give me all protein functions of genes expressed in the 5th chromosome).

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 [] that is proven to scale up to 15 billion explicit statements with included inference materialization. Our methodology includes several steps:

1. Transform all existing data sources to RDF
2. Load the data into RDF warehouse and compute light-weight inference
3. Execute post-processing steps for linked data instance alignment

4. Generate two types of semantic annotations optimized for high-recall and high-precision retrieval

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 [], LODD [], 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 followed the database schema strictly. 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 ommitted 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 [7]. Table X presents all datasets included PIKB that available for download from the LLD website.

Table 1. Pathway and Interaction Knowledge Base data sources.

Data source	Statements (explicit)	Schema	Description
Uniprot	1,146,084,021	Supplied by the provider	Protein sequences and annotations
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.

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.

[1] The dataset contains duplicated statements.

[2] Modified version to remove cycles in the hierarchy.

3.1 URI Naming Conventions

Linked Data principles state that for every URI it must be possible to deference it and access the related meta-data, [4]. 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 are the databases distributed into BioPAX format. We consider the cross tool interoperability more important thus we stick to 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: http://linkedlifedata.com/resource/db/type/id
R3. Identify the graph names with http://linkedlifedata.com/resource/db
R4. Name all generated predicate URIs http://linkedlifedata.com/resource/db/predicate
R5. Generate stable new URIs based on unique label to describe the resource (see dataset provenance and updates)

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 constrains 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 / Linking Linked Data

Extraction, transformation, loading (ETL) is a typical phase of 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 X 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 subset selection and the rule application is controlled by human.

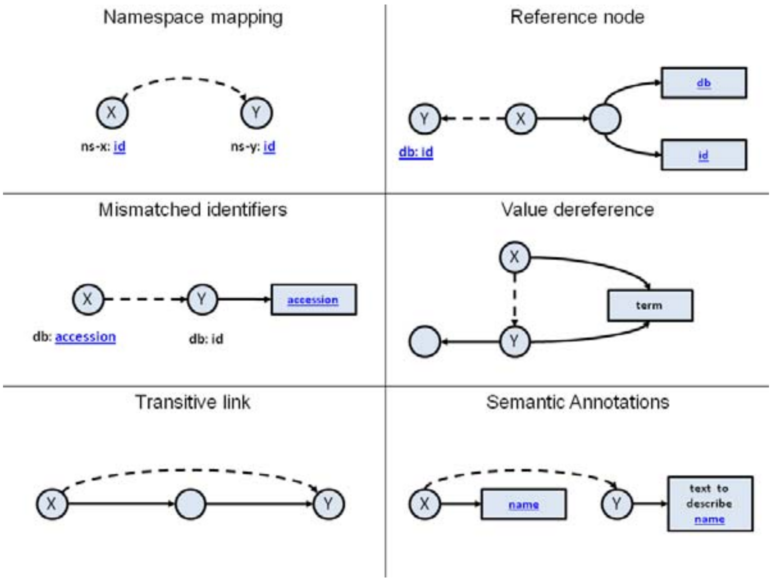

Figure 1. Linked Data instance alignment rules.

3.4 Semantic Annotation Linking

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 [5], a promising community effort to transform Wikipedia. In the DBPedia data the predicate *wikilink* connects resources representing linked pages in Wikipedia. This type of connection shares many similarities with the idea of semantic annotation [6] and Wikipedia could be referred as the biggest source of human curated semantic annotations. In LLD we use semantic annotation to add additional links between resources and to demonstrate the excellent multibillion statements scalability of the OWLIM repository.

4 Results and Discussions

With no significant optimizations the LLD prototype and the underlying OWLIM engine on a server configuration: 2 x E5420 @ 2.50GHz (4HT) 2GHz 64GB memory 900GB SAS RAID-5 8xSEAGATE Cheetah 15k 146GB HDD. It seems 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. Their processing is a realized with a custom Java rule language configured with set of SPARQL queries and function primitives. Table 3 presents statistics for the generated links between the data sources.

Semantic annotations are generated on the basis of a pipeline of text processing components, e.g., a semantic gazetteer and rule based filters of gazetteer matches. We use the semantic gazetteer to look-up in the text for predefined strings out of predefined lists. The lists are generated on the basis of a part of the LLD data and the strings, e.g. the instance names follow a particular schema, based on UMLS, e.g., the text is annotated against an ontology by giving each instance found in the text a class and instance identifier (i.e. unified resource identifier - URI). The recognized instances are subsequently imported back to the LLD database.

The semantic gazetteer lists are generated by selecting all names of particular instances in LLD and compiling a list of 2,518,641 strings. Since we found that some instances share common aliases (alternative names), as a second experiment we compiled a stricter list set of 1,248,890 strings. We selected more than 16 million strings, containing ~ 15 million strings from PubMed, as text for annotation. The extraction algorithm generated 705,338,334 semantic annotations while the stricter approach generated 263,323,164.

The introduced data integration approach demonstrated that it is highly efficient in the process of interlinking of highly heterogeneous data. The conventional semantic integration, supported by custom linked data mapping rules and semantic annotations provided us the possibility to ask questions with very high level of complexity, which requires the utilization of knowledge from diverse domain – sequence properties, molecular data and interactions, pharmacological information and clinics.

For the evaluation of the proposed approach we have developed a step-by step use case, which demonstrates the potential of the technology to provide more holistic view over a particular scientific problem, providing valuable insights from much wider perspective.

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 6”) which expands step by step the knowledge and at the same time increase the specificity of the asked question.

The queries follow the following pattern:

“Select all human genes (query 1), which code for proteins with known molecular interactions (query 2) and are analyzed with molecular techniques (query 3) like 'Epitope Mapping' (query 4). We can go even further in the mining as we can restrict the results just to gene/proteins, which are known drug targets (query 5) for a specific disease (query 6).”

Table 3. Linked Data mapping rules output

Source dataset	Destination dataset	Linked Data Mapping Rule	Number of connections	Semantic relationship
All BioPAX schema datasets	Entrez Gene	Reference Node	7,897	skos:closeMatch
All BioPAX schema datasets	Gene Ontology (UMLS)	Reference Node	44,642	skos:relatedMatch
All BioPAX schema datasets	NCBI Taxonomy (UMLS)	Reference Node	52,851	skos:closeMatch
All BioPAX schema datasets	UniProt knowledge base	Reference Node	107,183	skos:exactMatch
Diseasome (LODD)	Entrez Gene	Mismatched Identifiers	2,772	skos:closeMatch

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 in 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, UMLS for making them publicly available.

References

1. 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
2. Ben Vandervalk; Luke McCarthy; Mark Wilkinson, CardioSHARE: Web Services for the Semantic Web, Semantic Web Challenge 2008, 2008
3. Health Care and Life Sciences Interest Group: A Prototype Knowledge Base for the Life Sciences, http://www.w3.org/TR/hcls-kb/
4. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
5. 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 .

6. Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, Damyan Ognyanoff , Elsevier's Journal of Web Semantics, Vol. 2, Issue (1), 2005.
7. 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