



LarKC

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

FP7 – 215535

---

## **D7a.3.1 Prototype v1**

---

**Coordinator: Vassil Momtchev**

**With contributions from: Deyan Peychev, Todor Primov,  
Georgi Georgiev, Rostislav Hristov, ONTO; Bo  
Andersson, AZ**

**Quality Assessor: Angus Roberts  
Quality Controller: Vassil Momtchev**

Document Identifier:	LarKC/2008/D7a.3.1 /v1.0
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	1.0
Date:	30.09.2009
State:	Final
Distribution:	Public



## EXECUTIVE SUMMARY

The report is supplementary to the M18 software prototype deliverables. Since the initial M6 prototype, we have switched to a much bigger list of datasets and knowledge base, which includes more than 20 different data sources. Rule based and information extraction techniques are applied to the loaded RDF resources to increase the network interconnection and to capture the semantic relationship interpretations of each individual data source. The data generation workflow now has increased complexity, finer detailed steps, and replaces the initial four pipelines. In addition to the information transformation and the LLD loading process, a new front-end to the knowledge base is developed. It aims to assist advanced users in profiling, analyzing, and verifying the information and to be used for access by other software components. Each result screen also supports computer-friendly data formats that can be negotiated with 303-HTTP content accepted headers. Depending on the page content, they support RDF, N3, Turtle, JSON, SPARQL XML formats.



## DOCUMENT INFORMATION

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

<b>Deliverable</b>	<b>Number</b>	D7a.3.1	<b>Title</b>	Prototype v1
<b>Work Package</b>	<b>Number</b>	WP7a	<b>Title</b>	Semantic Integration for Early Clinical Development

<b>Date of Delivery</b>	<b>Contractual</b>	M18	<b>Actual</b>	M18
<b>Status</b>	version 1.0		final	<input type="checkbox"/>
<b>Nature</b>	prototype <input checked="" type="checkbox"/> report <input type="checkbox"/> dissemination <input type="checkbox"/>			
<b>Dissemination level</b>	public <input checked="" type="checkbox"/> consortium <input type="checkbox"/>			



<b>Authors (Partner)</b>	Vassil Momtchev, Deyan Peychev, Todor Primov, Georgi Georgiev, Rostislav Hristov (Ontotext); Bo Andersson (AstraZeneca)			
<b>Responsible Author</b>	<b>Name</b>	Vassil Momtchev	<b>E-mail</b>	vassil.momtchev@ontotext.com
	<b>Partner</b>	ONTO	<b>Phone</b>	

<b>Abstract (for dissemination)</b>	The report is supplementary to the M18 software prototype deliverables. Since the initial M6 prototype, we have switched to a much bigger list of datasets and knowledge base, which includes more than 20 different data sources. Rule based and information extraction techniques are applied to the loaded RDF resources to increase the network interconnection and to capture the semantic relationship interpretations of each individual data source. The data generation workflow now has increased complexity, finer detailed steps, and replaces the initial four pipelines. In addition to the information transformation and the LLD loading process, a new front-end to the knowledge base is developed. It aims to assist advanced users in profiling, analyzing, and verifying the information and to be used for access by other software components. Each result screen also supports computer-friendly data formats that can be negotiated with 303-HTTP content accepted headers. Depending on the page content, they support RDF, N3, Turtle, JSON, SPARQL XML formats.
<b>Keywords</b>	Linked Data, RDF warehouse, data integration, life sciences, early clinical development

<b>Version Log</b>			
<b>Issue Date</b>	<b>Rev. No.</b>	<b>Author</b>	<b>Change</b>
1/08/2009	0.1	Vassil Momtchev	Initial ToC
1/09/2009	0.2	Vassil Momtchev	Draft
20/09/2009	0.3	Vassil Momtchev	Changes in the workflow version
28/09/2009	0.4	Vassil Momtchev	Pre-final version send for QA
28/09/2009	0.5	Gergana Petkova	Proof-reading
29/09/2009	1.0	Vassil Momtchev	Implemented QA comments



## 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



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



## **TABLE OF CONTENTS**

<b>LIST OF FIGURES .....</b>	<b>7</b>
<b>LIST OF TABLES .....</b>	<b>8</b>
<b>LIST OF ACRONYMS.....</b>	<b>9</b>
<b>1. INTRODUCTION .....</b>	<b>10</b>
<b>2. PIKB AND OTHER RELEVANT DATASETS .....</b>	<b>12</b>
<b>3. DEVELOPED PLUG-INS .....</b>	<b>14</b>
3.1. OBO2SKOS.....	14
3.2. RDBMS2RDF.....	15
3.3. RDF SYNTAX VALIDATION.....	15
3.4. ONTOLOGY INSTANCE ALIGNMENT.....	15
3.5. PARALLEL SEMANTIC ANNOTATOR.....	17
<b>4. SEMANTIC DATA INTEGRATION WORKFLOW .....</b>	<b>18</b>
<b>5. FRONT-END AND EXTERNAL INTERFACES.....</b>	<b>19</b>
<b>6. CONCLUSION .....</b>	<b>20</b>
<b>7. REFERENCES .....</b>	<b>21</b>



## List of Figures

Figure 1 The pyramid of technologies used in the pharmaceutical R&D industry .....	10
Figure 2 Gene Ontology represented in OBO format .....	14
Figure 3 The minimal input for RDF syntax validation plug-in.....	15
Figure 4 LLD mapping rules to align ontology instances .....	17
Figure 5 Semantic data integration workflow. ....	18
Figure 6 LLD front-end pages and possible transitions .....	19



## List of Tables

Table 1 The list of datasets and their size expressed in RDF statement count.....	12
Table 2 Linked Open Drug Data datasets.....	13
Table 3 Literature references and thesauri datasets.....	13



## List of Acronyms

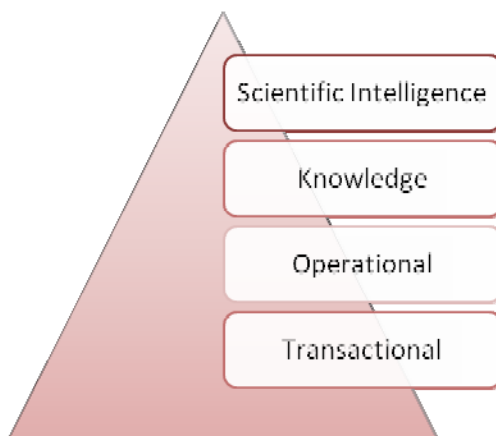
<b>Acronym</b>	<b>Description</b>
API	Application Programming Interface
ECD	Early Clinical Development
ETL	Extraction Transformation Loading (a typical data warehouse process)
GO	Gene Ontology
KB	Knowledge Base
LarKC	Large Knowledge Collider
LLD	Linked Life Data
OBO	Open Biomedical Ontologies
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
PIKB	Pathway and Interaction Knowledge Base
RDF	Resource Descriptor Framework
RDFS	RDF Schema
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
UMLS	Unified Medical Language System
KOS	Knowledge Organisation System
SKOS	Simple Knowledge Organisation System

## 1. Introduction

This deliverable report is supplementary to the Linked Life Data (LLD) software prototype running at [2]. The document describes the first prototype of WP7a “Semantic Integration for Early Clinical Development” use case [4], which is implemented on top of the LarKC platform, [1]. The scope of the current deliverable is limited to the documentation of the developed software plug-ins, the workflow of generating the knowledge base, and the front-end of the system. A detailed description and statistics of the developed dataset and the mappings to other data sources are published in the deliverable D7a.2.1 “Pathway and Interaction Knowledge Base”, [3].

LLD is a semantic data integration solution that simplifies the interlinking and the complex query evaluation of related biomedical data sources. Our approach to integrate multiple domain ontologies, biomedical thesauri, and databases is described in the literature as RDF warehousing [5], [6] or RDF mashup, [7]. An important difference from the described approaches is the use of formal reasoning to derive new implicit information. In the system analysis, we position the LLD prototype as some sort of meta-system, because it aggregates data from multiple independent information silos. Figure 1 shows the four levels of systems:

- **Transactional** level covers systems optimized for data storage. Typically, they perform very efficiently Online Transaction Processing (OLTP);
- **Operational** level is best described by specific laboratory experiment systems aimed to automate concrete tasks;
- **Knowledge** level is characterized by systems that Extract, Transform, and Load (ETL) information processes from various sources and generate a model that supports efficient online analytics processing (OLAP) – e.g. the information inference could be considered also as a special form of indexing;
- **Scientific Intelligence** level is a term that describes the business intelligence process applied to the scientific field. At this level we expect to run different algorithms to mine the knowledge and test new hypotheses;



**Figure 1 The pyramid of technologies used in the pharmaceutical R&D industry**

The scope of the LLD prototype is to investigate and address the challenges and the opportunities in the field of early clinical development, primarily at the level of knowledge and scientific intelligence.

Since the M6 proof-of-concept system described in [4], we have changed our work in two directions. Firstly, we have increased the number of used datasets and have found that some of the initially integrated ones do not provide sufficient quality or value. Secondly, the initial four LarKC platform pipelines have been replaced with a single workflow (e.g. the project switched to a more expressive and complex plug-in composition capabilities) responsible for performing all necessary steps for the knowledge base loading process.



Chapter 2 contains information about the used datasets and the type of processing needed to load them in the LLD.

Then, Chapter 3 presents the internals of the developed plug-ins and explains the type of tasks automated in the knowledge base creation process.

The LarKC workflow unifying and executing all plug-ins is presented in Chapter 4.

In the final Chapter 5 we present the current LLD front-end and the supported interaction interfaces.



## 2. PIKB and Other Relevant Datasets

The Pathway and Interaction Knowledge Base is a dataset, developed in the context of WP7a, that describes genes, proteins, interactions, and metabolic pathways. Its objective is to model information for cellular processes on the molecular level. Still, due to the limitations of today’s biological science, this knowledge is insufficient for fully understanding the specimen on the organism or even the physiological level. Hence, the researchers would like to understand the relationships between: genes, proteins, pathways, targets, diseases, drugs, and patients. Each instance of these types has a rich meta-data describing its specificity. Ultimately, scientists need to put into context the existing or experimentally proven information – e.g. to list all drugs, related to asthma, that are known to be related to pathways or interactions, involved into inflammatory responses. Despite the fact that, when multiple biomedical data sources are integrated, the answer is readily available, liking and accessing this knowledge may be a difficult and time-consuming task for the researchers. PIKB integrates major public biomedical sources describing molecular interactions, genes, and proteins and interlinks the information.

Table 1 presents a list of identified biomedical data sources that contain valuable information for understanding the molecular interaction processes. Every data source is converted to RDF if not already distributed in that format. The BioPAX ontology format distributed by Pathway Commons database is now used as a primary source for interactions and pathways.

<b>Data source</b>	<b>Statements (explicit)</b>	<b>Schema</b>	<b>Description</b>
Uniprot	1,146,084,021	Supplied by the provider	Protein sequences and annotations
Entrez-Gene	107,193,308	Custom schema	Genes and annotation
BioGRID	9,454,917	BioPAX 2.0 (distributed by Pathway Commons)	Protein interactions extracted from the literature
Cell Map	151,788	BioPAX 2.0 (distributed by Pathway Commons)	Cancer related signaling pathways
HPRD	1,805,651	BioPAX 2.0 (distributed by Pathway Commons)	Information on human protein functions
IMID	154,408	BioPAX 2.0 (distributed by Pathway Commons)	General Repository for Interaction Datasets
NCI Nature	454,876	BioPAX 2.0 (distributed by Pathway Commons)	Pathway and interaction database
MINT	7,915,613	BioPAX 2.0 (distributed by Pathway Commons)	Molecular interaction database
Reactome	839,084	BioPAX 2.0 (distributed by Pathway Commons)	Human pathways and interactions
IntAct	11,005,555	BioPAX 2.0 (distributed by Pathway Commons)	Protein interaction database
Total	1,285,059,221	-	All PIKB data sources

**Table 1 The list of datasets and their size expressed in RDF statement count**

The RDF format and Linked Data principles allow easy consolidation of the datasets made by other members of the community. The Health Care and Life Science Interest Group is a community group, part of the W3C, [9]. The Linked Open Drug Data (LODD) taskforce [8] is dedicated to collect data that describes drug related information. The LODD is considered relevant to WP7a “Semantic Integration for Early Clinical Development” and therefore it is interconnected with the PIKB dataset. We created relationships that link semantically related information and identify the redundant concepts. Table 2 contains the size of LODD and other related datasets.



Data source	Statements (explicit)	Schema	Description
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 the provider	ClinicalTrials.gov represented into RDF
DBPedia <sup>1</sup>	439,775,096	Supplied by the provider	Structured information from Wikipedia
Total	447,587,674	-	All LODD dataset

**Table 2 Linked Open Drug Data datasets**

The WP7a warehouse is also extended with unstructured information from scientific journals. The UMLS meta-thesaurus and the OBO ontologies are another knowledge dimension that formally defines a large number of biomedical and health related concepts within a light-weight ontological representation format. Their main purpose is the generation of semantic annotations in the text field of the other data sources and future query disambiguation. Table 3 contains additional data sources that describe common biomedical knowledge in the form of thesauri or unstructured text.

Data source	Statements (explicit)	Schema	Description
PubMed	807,851,455 <sup>2</sup>	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 <sup>3</sup>	12,420,882	SKOS	Database that contains information about biomedical and health related concepts, their various names, and the relationships among them
Human Phenotype Ontology	70,911	SKOS	Human phenotype ontology
Symptom Ontology	4,163	SKOS	Symptoms ontology
Disease Ontology	446,066	SKOS	Diseases Ontology
Total	820,794,845	-	All data sources

**Table 3 Literature references and thesauri datasets**

<sup>1</sup> Modified version to remove cycles in the hierarchy

<sup>2</sup> The dataset contains duplicated statements

<sup>3</sup> Composed by data sources with UMLS license restriction – Category 1 and 2



### 3. Developed Plug-ins

The LarKC platform provides a plug-in based framework that allows the incorporation of multiple components into a complex execution workflow. To implement the batch loading and the transformation steps of “Semantic Data Integration Workflow” (see chapter 4), a sequence of Transformers is used.

#### 3.1. OBO2SKOS

The OBO flat file format is an ontology representation language. According to its authors, it is designed to represent a subset of the concepts in the OWL description logic language, with several extensions for meta-data modelling and the modelling of concepts that are not supported in description logic languages, [12]. Figure 2 presents a part of the Gene Ontology expressed in OBO syntax.

```
[Term]
id: GO:0000001
name: mitochondrion inheritance
namespace: biological_process
def: "The distribution of mitochondria, including the mitochondrial genome,
into daughter cells after mitosis or meiosis, mediated by interactions
between mitochondria and the cytoskeleton." [GOC:mcc, PMID:10873824,
PMID:11389764]
synonym: "mitochondrial inheritance" EXACT []
is_a: GO:0048308 ! organelle inheritance
is_a: GO:0048311 ! mitochondrion distribution

[Term]
id: GO:0000002
name: mitochondrial genome maintenance
namespace: biological_process
def: "The maintenance of the structure and integrity of the mitochondrial
genome; includes replication and segregation of the mitochondrial
chromosome." [GOC:ai, GOC:vw]
is_a: GO:0007005 ! mitochondrion organization

[Term]
id: GO:0000003
name: reproduction
namespace: biological_process
alt_id: GO:0019952
alt_id: GO:0050876
def: "The production by an organism of new individuals that contain some
portion of their genetic material inherited from that organism."
[GOC:go_curators, GOC:isa_complete, ISBN:0198506732 "Oxford Dictionary of
Biochemistry and Molecular Biology"]
subset: goslim_generic
subset: goslim_pir
subset: goslim_plant
subset: gosubset_prok
synonym: "reproductive physiological process" EXACT []
xref: Wikipedia:Reproduction
is_a: GO:0008150 ! biological_process
```

**Figure 2 Gene Ontology represented in OBO format**

Today, there are more than 89 ontologies published at the OBO Foundry website, [13] in OBO format. The lack of RDF representation makes the ontologies inaccessible to LLD and other Semantic Web applications. The formal OBO semantics are defined in [11]. However, considering the scale of LLD, a lightweight approach may be more appropriate. [9] suggests using simpler knowledge



organisation system (SKOS) to represent the information in a tractable logic fragment. The simple knowledge organisation system (SKOS) vocabulary is used to map most of the OBO format properties to RDF. We used the specification described in [9] to implement a transformation tool. The work is packaged as LarKC Transformer plug-in.

### 3.2. RDBMS2RDF

A large number of biomedical data sources are distributed in the form of relational database dumps or export scripts. Their integration in LLD is a two phase process. First, a database server that supports the distributed dump or script format is setup. Then, RDBMS2RDF is used to output the data into RDFXML format.

The plug-in encapsulates a component part of the ORDI framework that realises a practical and efficient way to interact with information, stored in relational databases. The mapping of the database columns to predicate is described by an ontology. The plug-in receives as input the output location and the directory that contains the transformation descriptions.

### 3.3. RDF Syntax Validation

This is a light-weight validation plug-in that checks the correctness of the generated data. It also supports statement counting and separation of sameAs statements (e.g. in specific data layer configurations the initial sameAs statement import reduces the inference complexity). Figure 3 shows the plug-in input and output.

```
Input:

@prefix lld: <http://linkedlifedata.com/resource/>.
@prefix importer: <http://linkedlifedata.com/resource/importer#>.

importer:Location importer:defaultGraph lld:uniprot ;
  importer:defaultNS <http://purl.uniprot.org/uniprot/> ;
  importer:URL < file:///C:/datasource/uniprot>.

Output:

@prefix lld: <http://linkedlifedata.com/resource/>.
@prefix importer: <http://linkedlifedata.com/resource/importer#>.
@prefix uniprot: <file:///C:/datasource/uniprot#>.

importer:Location importer:defaultGraph lld:uniprot ;
  importer:defaultNS < http://purl.uniprot.org/uniprot/>;
  importer:URL < file:///C:/datasource/uniprot>.

uniprot:uniprot.rdf importer:statementsCount "34644625" ;
  importer:sameAsCount "120" .
```

**Figure 3 The minimal input for RDF syntax validation plug-in.**

### 3.4. Ontology Instance Alignment

The warehousing of multiple RDF data sources generated by different authors has lead to inconsistencies in the conceptual model. Some of the concepts expressed in the RDF model remain disconnected despite being semantically related. In the course of the semantic data integration process we identified six patterns of malformed generated data.

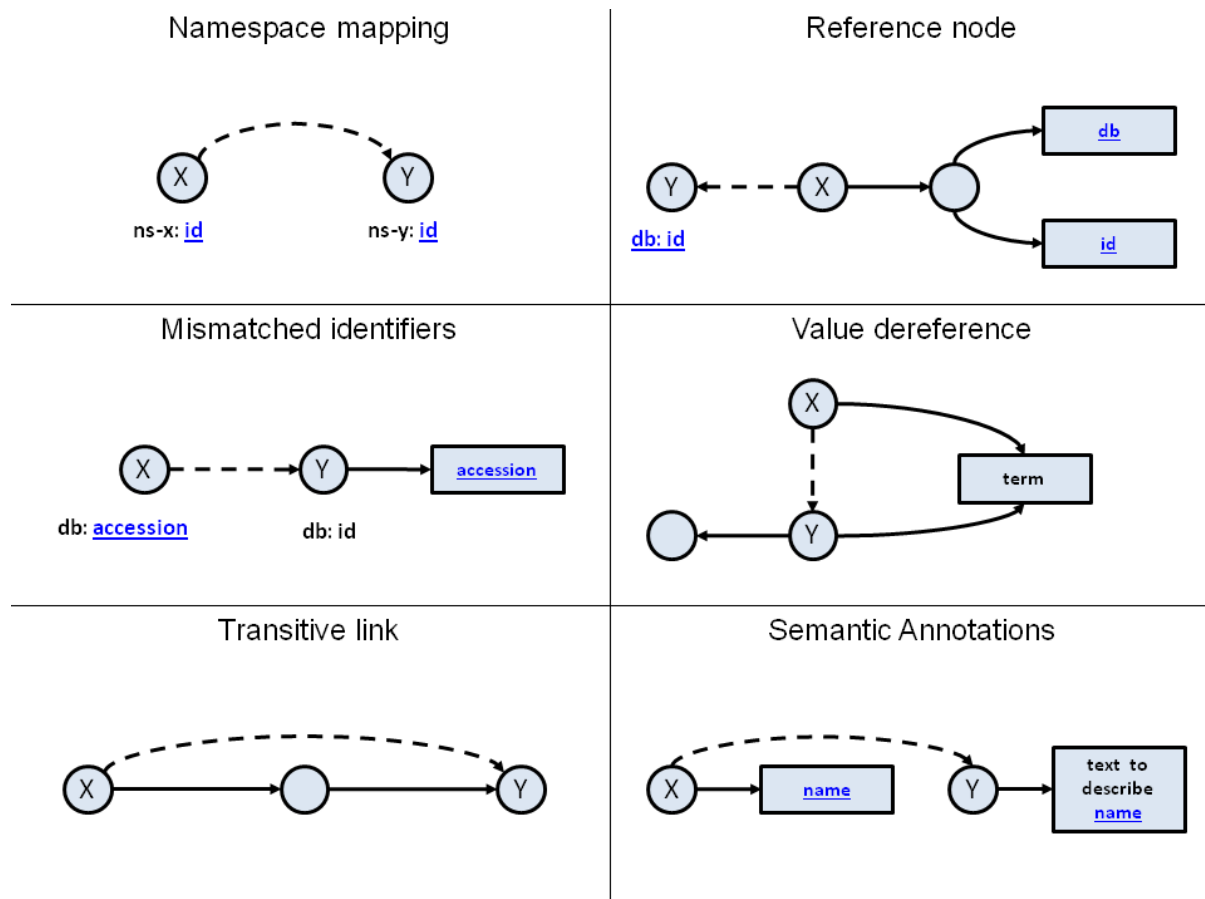
1. Namespace mapping - Two RDF datasets use one and the same local identifier but define different namespaces; this case is typical for data sources distributed in RDF format that refer



to common database identifiers (like GO, Entrez-Gene, etc.) without resolvable URI supported by the authors.

2. Reference node - A trick to prevent the former pattern that uses a reference dummy node to designate the database id and name; this is also the recommended way to create cross-references to external data sources in the BioPAX specification. However, the nodes remain disconnected even after the data sources were imported.
3. Mismatched identifier - Database entries have multiple identifiers used for different purposes. For example, the Entrez-Gene database has a gene symbol (alphanumeric string) and an id (numeric value). The id stands for a composite key constituted by a unique combination of gene symbol and organism.
4. Value dereference - A lazy-way to reference controlled vocabularies by using only the concept name, but not the identifier; PubMed is annotated with the name of the MeSH term names, but not MeSH concept ids.
5. Transitive link - A pattern used to link two data sources based on the common relation to a third-one; DBPedia has links to Freebase and ICD-10 codes.
6. Semantic annotation - A pattern used to link resources based on literals that enumerate a list of named entities; DrugBank indication field lists a sequence of diseases.

Figure 4 presents the six patterns pictographically. The red lines express the existing explicit relationships that are used to map the data. The dashed lines and the underlined text of the captions (e.g. used either as part of the URI or literals) designate the criteria for mapping the information. The specified mapping rules are not universally applicable for all RDF types. They are not easily expressible into formal languages such as SPARQL [14] or R-entailment rules [15], either. For example, rules 1, 2, 3 heavily depend on the efficient URI decomposition and string matching functions. Rule 6, if applied to highly inflection types like genes or proteins name recognition, is a result of complex natural language processing.



**Figure 4 LLD mapping rules to align ontology instances**

From the LarKC plug-in view point this is a reasoning process that performs domain specific ontology instance alignment of concrete datasets. From the warehousing view point the connections reduce the complexity of ETLs to prepare the data. The mapping rules are defined by pairs of SPARQL queries that return the RDF resources to be selected as a relationship’s source and destination.

### 3.5. Parallel Semantic Annotator

The Semantic Annotation Transformer plug-in was developed in the context of WP2 “Selection and Retrieval”, [16]. Due to the limitations of the LarKC platform, the path to the document that will be analyzed is hardcoded. In the context of WP7a, a new extended version is implemented and referred to as Parallel Semantic Annotator. It features:

- Multi-core parallel information extraction – it supports multiple parallel GAP pipeline executions.
- Convention to store the document as part of a data layer – this removes the requirement to hardcode the document location.
- Convention to represent the semantic annotation as document annotation – it ensures a better isolation between the plug-in and the Gate Application logic.



## 4. Semantic Data Integration Workflow

LLD is a knowledge base composed by more than 20 different data sources. The workflow is used to automate the transformation steps of loading, mapping, and linking the information. Conceptually, it has to execute the following sequence of tasks. The responsible plug-in for each step is indicated as follows:

1. Transform the data to RDF (OBO2SKOS, RDBMS2RDF) – this is the initial phase, transforming all data formats to RDF; the step simplifies the next steps of information management and eliminates all data syntax differences.
2. Filter RDF statements or re-write predicates (RDF Transformer) – the phase allows simple RDF node transformation like alter predicate URIs; replace namespaces; or generate URI for blank nodes.
3. Validate that the RDF data is correctly generated and output statistics for the imported datasets (RDF Syntax Validation) – this is a complementary step aiming at early problem detection (e.g. incorrectly generated RDF syntax).
4. Import the data into the data layer (Importer) – typically, the RDF data import process is the most time and resources consuming task; when the data layer inference is enabled, this phase also materializes all implicit statements.
5. Execute ontology instance alignment (Ontology Instance Alignment) – the phase resolves the structural differences in the way the data is modelled into the RDF graph; the semantic ambiguity is also decreased.
6. Generate additional relationships with information extraction (Parallel Semantic Annotator) – the final step in the LLD generation process is the execution of a domain specific information extraction process that generates either new relationships or entities; the phase captures the specific semantics of the data and enables the transparent integration of relationships generated from structured and unstructured data sources.

Figure 5 presents conceptually the workflow of the Transformer that completes the process. Many of the workflow steps can be executed in parallel. In the current workflow implementation, the control-flow is managed by a scripted decider, where only the Ontology Instance Alignment and Importer are executed sequentially because of the internal specificity. The Parallel Semantic Annotator has a limited parallelisation capability, depending on the annotation pipeline specificity, and it is applicable only to multi-processor or -core environments.

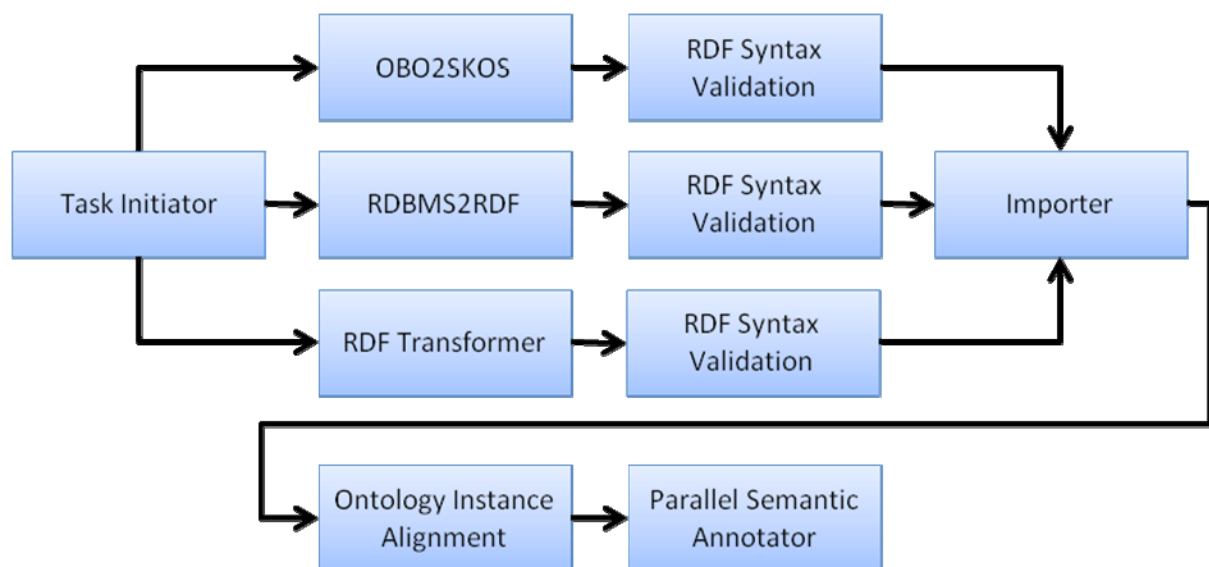


Figure 5 Semantic data integration workflow.

## 5. Front-end and External Interfaces

LLD is a semantic data integration platform that hosts large knowledge bases. Still, a major obstacle for the end-user is the possibility to interact with the existing information. As part of the M18 prototype, we have developed a base-line front-end prototype that guarantees a minimal set of supported presentation functionality. Figure 6 shows a diagram of the available screens and the possible transition between them.

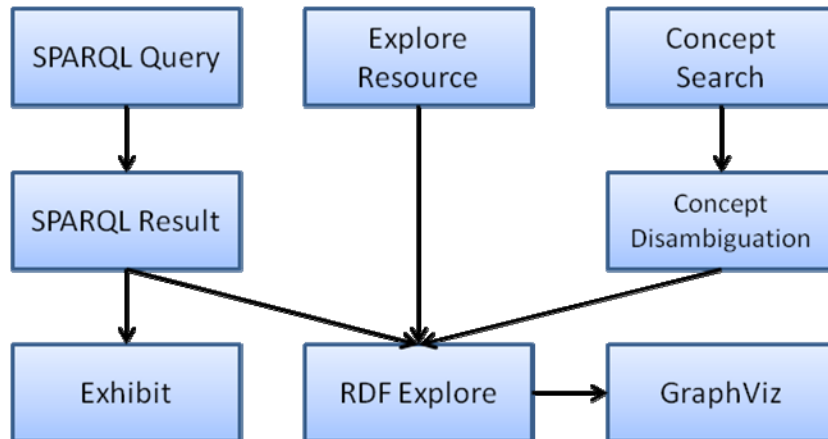


Figure 6 LLD front-end pages and possible transitions

LLD 0.3.4 has the following features:

- Three different searches
  - SPARQL Query – arbitrary SPARQL query
  - Explorer Resource – resolves a specific URI
  - Concept Search – full-text search in the concept literals
- Three different result visualisations
  - Exhibit – faceted search over RDF data
  - RDF Explorer – URI browsing
  - GraphViz – generates a graph presentation of the RDF model
- Each result screen supports the retrieval of displayed information into multiple data formats with 303-content negotiation; depending on the page content, the RDF, N3, Turtle, JSON, SPARQL XML formats are supported.

For the latest details, please refer to the LLD website<sup>4</sup>.

<sup>4</sup> <http://linkedlifedata.com>



## 6. Conclusion

This report presents the M18 status and the first prototype version of the WP7a “Semantic Integration for Early Clinical Development”. It provides a brief introduction to the used datasets, their schema and size expressed in RDF statements. Then, it presents the newly developed plug-ins in the context of WP7a use case: 1) OBO2SKOS – transforms OBO ontologies to RDF and SKOS vocabulary, 2) RDBMS2RDF – wraps the ORDI framework and the RDBMs to RDF translation, 3) RDF Syntax Validation – helps in the statistics generation, 4) Ontology Instance Alignment – aligns semantically related instances, and finally 5) Parallel Semantic Annotator – links the unstructured textual descriptions with the ontology instances. The WP7a data integration and transformer workflow is composed by the listed plug-ins and executed by a scripted decider.

Finally, a brief description of the present knowledge base front-end is included. For the latest releases and information updates, please refer to the LarKC [1] and LLD [2] websites.



## 7. References

- [1] <http://www.larkc.eu>
- [2] <http://linkedlifedata.com>
- [3] Momtchev, V. et al., PIKB, LarKC project deliverable D7a.2.1, 2009
- [4] Anderson, B., Momtchev V., Requirements summary and data repository, LarKC project deliverable D7a.1.1, 2008 .
- [5] Cheung, Kei-Hoi; Yip, Kevin Y; Smith, Andrew; Deknikker, Remko; Masiar, Andy & Gerstein, Mark: YeastHub: a semantic web use case for integrating data in the life sciences domain. In: *Bioinformatics* , Vol. 21 Suppl 1 , June (2005) , S. i85-i96 .
- [6] Erling, Orri & Mikhailov, Ivan: *RDF Support in the Virtuoso DBMS.* , Vol. 113 GI (2007) , S. 59-68 .
- [7] 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 .
- [8] Linked Open Drug Data website – <http://esw.w3.org/topic/HCLSIG/LODD>
- [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] World Wide Web Consortium (W3C) website – <http://www.w3.org/>
- [11] Horrocks, I., OBO Flat File Format Syntax and Semantics and Mapping to OWL Web Ontology Language .
- [12] Day-Richter, J., The OBO Flat File Format Specification, version 1.2 – [http://www.geneontology.org/GO.format.obo-1\\_2.shtml](http://www.geneontology.org/GO.format.obo-1_2.shtml)
- [13] OBO Foundry website - <http://www.obofoundry.org/>
- [14] Prud'hommeaux, E., and Seaborne A., SPARQL Query Language for RDF W3C Recommendation, 15 January 2008 .
- [15] ter Horst, HJ: Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary In: *Web Semantics: Science, Services and Agents on the World Wide Web* , Vol. 3 , Nr. 2-3 (2005) , S. 79-115 .
- [16] Cunningham, H., D2.2.1, 2.5.1 Month 12 Selection Components (report accompanying two software deliverables) LarKC project deliverable, 2009 .