



## LarKC

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

FP7 – 215535

---

# D7a.3.3 Prototype v3

---

**Coordinator: Vassil Momtchev**

**With contributions from: Deyan Peychev, Konstantin  
Pentchev, Todor Primov, Bo Andersson**

**Quality Assessor: Mark Greenwood  
Quality Controller: Vassil Momtchev**

Document Identifier:	LarKC/2010/D7a.3.3 /v1.0
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	1.0
Date:	29.09.2011
State:	final
Distribution:	public



## EXECUTIVE SUMMARY

This deliverable briefly describes the final prototype version of Linked Life Data (LLD) and the application of the LarKC platform as new exploration methods for challenges such as drug discovery, genetic epidemiology of cancer and other diseases, and a carcinogenesis research. LLD aggregates more than 27 public domain databases and collects over 5 billion facts, describing all types of biomedical knowledge. The service offers a solid dataset for research and experiments of various reasoning and information extraction techniques, and at the same time applies a new type of knowledge representation methods and infrastructure to help better analyze the information relevant for the drug development process.



## 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>	<a href="http://www.larkc.eu/">http://www.larkc.eu/</a>		
<b>Document URL</b>			
<b>EU Project Officer</b>	Stefano Bertolo		

<b>Deliverable</b>	<b>Number</b>	D7a.3.3	<b>Title</b>	Prototype v3
<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>	M 42	<b>Actual</b>	
<b>Status</b>	version 1.0		final x	
<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, Konstantin Pentchev, Todor Primov (Ontotext), Bo Andersson (AstraZeneca)			
<b>Responsible Author</b>	<b>Name</b>	Vassil Momtchev	<b>E-mail</b>	vassil.momtchev@ontotext.com
	<b>Partner</b>	Ontotext	<b>Phone</b>	

<b>Abstract (for dissemination)</b>	The document presents the final update of Linked Life Data (LLD) platform in the context of the LarKC project. LLD is the main deliverable of WP7a "Semantic Integration for Early Clinical Development". In this work package the LarKC platform technology is used to support the development of new exploration methods for challenges such as: drug discovery, genetic epidemiology of cancer and a carcinogenesis research by providing a novel semantic-enabled view towards the existing information. Currently, the LLD service is publicly operational at <a href="http://linkedlifedata.com">http://linkedlifedata.com</a> . A production service is also deployed for internal research in the AstraZeneca intranet. Ontotext will continue to develop LLD service even beyond the LarKC project as directly exploitable resource.
<b>Keywords</b>	LLD, linked data, RDF, warehouse, drug development

<b>Version Log</b>			
<b>Issue Date</b>	<b>Issue Date</b>	<b>Issue Date</b>	<b>Issue Date</b>
20/12/2010	20/12/2010	20/12/2010	20/12/2010

## PROJECT CONSORTIUM INFORMATION

Participant's name	Partner	Contact
Semantic Technology Institute Innsbruck, Universitaet Innsbruck	 	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.		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		Volker Tresp, SIEMENS AKTIENGESELLSCHAFT, Muenchen, Germany, E-mail: volker.tresp@siemens.com
THE UNIVERSITY OF SHEFFIELD		Hamish Cunningham, THE UNIVERSITY OF SHEFFIELD Sheffield, UK, Email: h.cunningham@dcs.shef.ac.uk



<p>VRIJE UNIVERSITEIT AMSTERDAM</p>		<p>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>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 Cancer Centre International de Recherche sur le Cancer</p>	<p>Paul Brennan, INTERNATIONAL AGENCY FOR RESEARCH ON CANCER, Lyon, France, Email: brennan@iarc.fr</p>
<p>INFORMATION RETRIEVAL FACILITY</p>		<p>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. DATASET UPDATES.....</b>	<b>11</b>
<b>3. KNOWLEDGE BASE PROCESSING MECHANISM .....</b>	<b>12</b>
3.1. DATA STAGING.....	13
3.2. KNOWLEDGE BASE LOADING .....	13
3.3. KNOWLEDGE BASE MERGING.....	14
3.4. INSTANCE MAPPINGS .....	14
3.5. RELATION MAPPINGS .....	16
<b>4. CONCLUSION.....</b>	<b>19</b>
<b>REFERENCES.....</b>	<b>20</b>



## List of Figures

Figure 1 The number of entries in UniprotKB/TreMBL.....	12
Figure 2 OWLIM repository indexes .....	13
Figure 3 OWLIM Knowledge Base Loading Process .....	14
Figure 4 Patterns to align instance level identity over linked data.....	15
Figure 5 Extracting inclusion/exclusion criteria from between clinical trial and UMLS concepts.....	17



## List of Tables

Table 1 Data sources statistics.....	11
Table 2 The stage of LLD service .....	12
Table 3 Cross data source instance mapping statistics .....	16
Table 4 Relation mappings generated for the causality mining interface .....	18



## 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

The document presents the final update of the Linked Life Data (LLD) platform. LLD is the main deliverable of WP7a “Semantic Integration for Early Clinical Development”. In this work package the LarKC platform technology is used to support the development of new exploration methods for challenges such as: drug discovery, genetic epidemiology of cancer and carcinogenesis research by providing a novel semantic-enabled view towards the existing information. Currently, the LLD service is publicly operational at <http://linkedlifedata.com>. A production service is also deployed for internal research in the AstraZeneca intranet. Ontotext will continue to develop the LLD service even beyond the LarKC project as a directly exploitable resource.

The efficient data integration and interoperability is the biggest challenge faced by the existing technology. The ever increasing amounts of information generated by the distributed and decentralized data providers put a constant pressure on the data driven research of early clinical development. LLD aggregates 27 public domain databases and contains over 5 billion facts, describing all types of biomedical knowledge. The service offers a solid dataset for research and experiments of various reasoning and information extraction techniques, and explores a new type of knowledge representation methods and infrastructure to help better analyze the information relevant to the drug development process.

In the first three quarters of 2011 the public version of LLD served more than 21,000 unique visitors and showed an increased number of visits (returning visitors) compared with 2010 (17,500 unique visitors during the whole year). Its users generated more than 2.1 million hits that resulted in 20 GB network traffic (compared to only 5 GB for 2010). Until the end of the year we plan a 50% increase of the unique visitors and over 5 times more network traffic.

The final prototype version is focussed on delivering an optimized semantic data warehousing process of a large number of dynamic data sources and integrating advanced administrative tools for service administration. Chapter 2 presents an updated conceptual data integration methodology based on the RDF model and describes the latest updates and growth in the collected databases. Chapter 3 presents the revised RDF warehousing methodology. It lists the stages to complete the tasks of extract, transform and load data described by semantics, load it in the knowledge base, do incremental updates and finally performs instance and schema level alignment by scheduling jobs.



## 2. Dataset Updates

In the latest release (LLD 0.8) the number of the data sources has increased to 27, and the total number of statements is 5,120,886,447. There is a slight decrease in the number of statements since the last prototype report mainly due the removal of the DBPedia dataset (data quality problems, now it is replaced by Freebase) and the optimization in the PubMed schema that removed many artificially generated URIs.

Table 1 lists all current data sources, their loading date and the number of triples. The latest release table and dump files can be accessed from the LLD website [1].

Data source	Load date	Number of statements
BioGRID	10.12.10	12,660,756
CellMap	24.09.10	149,175
ChEBI	04.10.10	323,212
DailyMed	13.10.10	162,972
Disease Ontology	21.10.10	144,812
Diseasome	14.10.10	72,445
DrugBank	14.10.10	517,023
Freebase	15.04.10	395,958,356
Gene Ontology (GO)	01.03.11	320,182
HapMap	15.12.10	22,462,178
HPRD	24.09.10	1,961,200
Human Phenotype Ontology	13.10.10	84,378
HumanCYC	07.10.10	327,218
IMID	24.09.10	83,091
IntAct	03.02.11	16,669,066
LHGDN	08.03.10	316,020
LinkedCT	14.10.10	7,031,859
MetaCyc	07.10.10	1,709,326
MINT	21.12.10	21,353,848
NCBI Entrez-Gene	09.03.11	161,563,100
NCI Nature	24.09.10	610,689
PubMed	20.03.11	1,371,818,500
Reactome	15.03.11	814,807
SIDER	13.10.10	101,542
Symptom Ontology	15.10.10	4,163
UMLS	01.11.10	121,438,271
UniProt	01.03.11	2,354,085,964

**Table 1 Data sources statistics**

The Uniprot data source continues the steady growth of 28% compared to the prototype presented in [2]. On Figure 1 we can see the steady trend of doubling the number entries every two years. There are no other significant changes in the data sources schema the except the exclusion of DBPedia. The DBPedia dataset has significant overlap with Freebase and introduces extreme number of predicates (over 70K), which causes an increased memory foot-print due to the support of extra indexes.

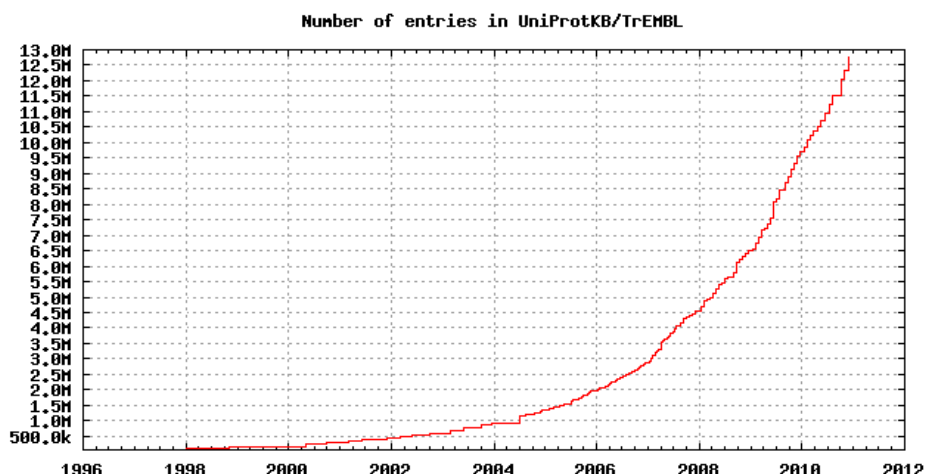


Figure 1 The number of entries in UniprotKB/TreMBL<sup>1</sup>

### 3. Knowledge Base Processing Mechanism

From LLD 0.7 version onwards we have implemented a new knowledge processing algorithm. It greatly reduces the time to perform administrative and maintenance activities such as modifying or excluding any number of datasets from the production service, keeping multiple versions of a single source and experimenting with multiple schemata alignment models. Table 2 presents a short overview of the LLD loading stages:

Stage Name	Input	Output	Description
Data staging	Data source distribution format	RDF data	Transforms all data in a common data format suitable for further processing
Knowledge Base Loading (new)	RDF data	OWLIM image for each data source	Loads all data into separate OWLIM images (i.e. the internal OWLIM binary format).
Knowledge Base Merging (new)	OWLIM images for each data source	Single OWLIM image	Merges a set of OWLIM images into a single repository
Instance mapping	LLD service + instance mapping rules	LLD service enriched with identity mapping statements	Aligns resources in the repository
Relation mapping (new)	LLD service + relation mapping rules	LLD service enriched with new schemata and relations between the resources	Introduces new schemata in the repository and generates new relations between the resources

Table 2 The stage of LLD service

<sup>1</sup> <http://www.ebi.ac.uk/uniprot/TrEMBLstats/>

### 3.1. Data staging

The data staging phase relies on a few simple and effective conventions necessary to resolve the syntactic level and data format heterogeneity problems. The following 6 basic conventions ensure that all original database identifiers are easily retrievable and compliant with the linked data principles:

1. Preserve the original RDF structure if distributed by the owner
2. Use resolvable URIs for the data sources with no RDF distribution
3. Construct the generated URIs in the form of `l1d:resource/db/type/id`
4. Identify the graph names with `l1d:resource/db`
5. Name all generated predicate URIs `l1d:resource/db/predicate`
6. Generate stable new URIs based on a unique label that describes the resource (see Dataset provenance and updates)

In a nutshell, we encourage all content providers to maintain resolvable URIs and govern the policy for resources publishing, update and retraction. In the scenarios where the URIs are not resolvable, LLD will act as a mediator that facilitates the end users in rewriting the URIs.

### 3.2. Knowledge Base Loading

The semantics of the corresponding RDFS and OWL properties `rdfs:seeAlso` and `owl:sameAs` is often misinterpreted and inconsistently applied across the different RDF datasets. The `rdfs:seeAlso` property does not directly result in any inferred statements. It poses a challenge only to the query developer who should consider its scope. However, the `owl:sameAs` property, if not properly used, produces a massive number of implicit statements and conceptual inconsistencies. For instance, when two concepts are merged in the data sources represented in SKOS, the statement of `owl:sameAs` is likely to result in multiple `skos:prefLabel` or `skos:inScheme` values. Another example from the LOD cloud is the assertion of an `owl:sameAs` statement between an organism specific gene and the abstract notion of a gene, where a full equivalence between all gene information will be automatically generated. Hence, the LLD loading process filters all `owl:sameAs` statements and replaces them with the less engaging and safer `skos:exactMatch` predicate. See the Instance Mappings section for more details about the instance level alignment.

Figure 2 presents the two main types of indexes maintained by the OWLIM repository. The Resource Index is a dictionary that translates all RDF resources (literals, URIs and blank nodes) to internal database identifiers for a more efficient processing. The Statement Index stores all statement information encoded with the internal resource identifiers mapped to the Resource Index.



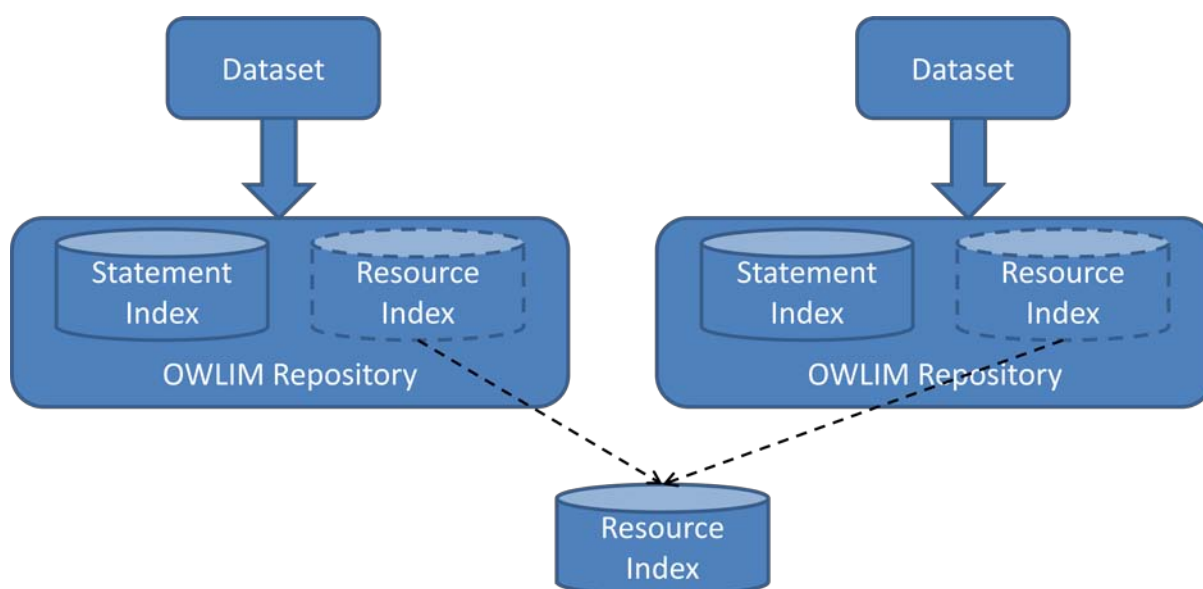
**Figure 2 OWLIM repository indexes**

A very distinctive feature of the LLD knowledge base loading process is that every dataset version is loaded into a dedicated repository, isolated from the other data OWLIM repository. Thus, the LLD data sources are federated into different independent repositories. The main benefit of federating the datasets is easier administration of the LLD service in terms of faster data loading, cheaper dataset bug fixing and simpler version control, enabling two independent release cycles. The cycle of the knowledge base loading is concerned only with the primary data source updates and it is executed

independently from the full LLD knowledge base release, performed during the Knowledge Base Merging stage.

### 3.3. Knowledge Base Merging

The main implication of using federated repositories is the extra complexity required in the SPARQL query optimizer and executor in order to retrieve information from multiple sources. Even if it seems like an excellent goal to pursue in the future, LLD 0.7 adheres to a more conservative hybrid approach of loading each knowledge base. All datasets are loaded sequentially in order to use a single Resource Index file, which contains mappings for the RDF resources in all data sources. Figure 3 represents the actual loading stages, which aim to reduce the complexity of merging any arbitrary number of preloaded datasets.



**Figure 3 OWLIM Knowledge Base Loading Process**

In essence the Knowledge Base Merging stage copies the global Resource Index (i.e. the dictionary index used to load all datasets) and merges all selected Statement Index files into a single file. The complexity of the past operation is pretty limited to the I/O file disk speed due the organisation of each Statement Index file. The whole process of selecting a certain number of datasets and producing a single merged repository takes several hours for the complete LLD service.

### 3.4. Instance Mappings

Once all data is loaded into the warehouse we ensure that the relevant resources are properly interlinked. The resource interlinking of the cross-data source guarantees that the redundant identifiers are associated with one of the three levels of relationship, reused from the SKOS schema [3]:

- <http://www.w3.org/2004/02/skos/core#exactMatch> – full resource equivalence that should be transitively propagated. It is used when there is a full overlap of the identifiers and the type of resources (e.g. Uniprot and BioPAX protein sequence identifier).
- <http://www.w3.org/2004/02/skos/core#closeMatch> – resource equivalence limited only to the information retrieval needs. It is used when there is an overall similarity but different context of the sources (e.g. BioPAX protein sequence and an identifier of the encoding gene).
- <http://www.w3.org/2004/02/skos/core#relatedMatch> – limited resource equivalence that facilitates the composition of complex queries without involving string matching functions.

All instance-level equivalence in LLD is derived by one of the patterns presented in Figure 2, i.e. after the replacement of an existing `owl:sameAs` predicate with `skos:exactMatch`. The solid lines express the existing explicit relationships 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 applicable in all cases, and are designed for specific datasets only.

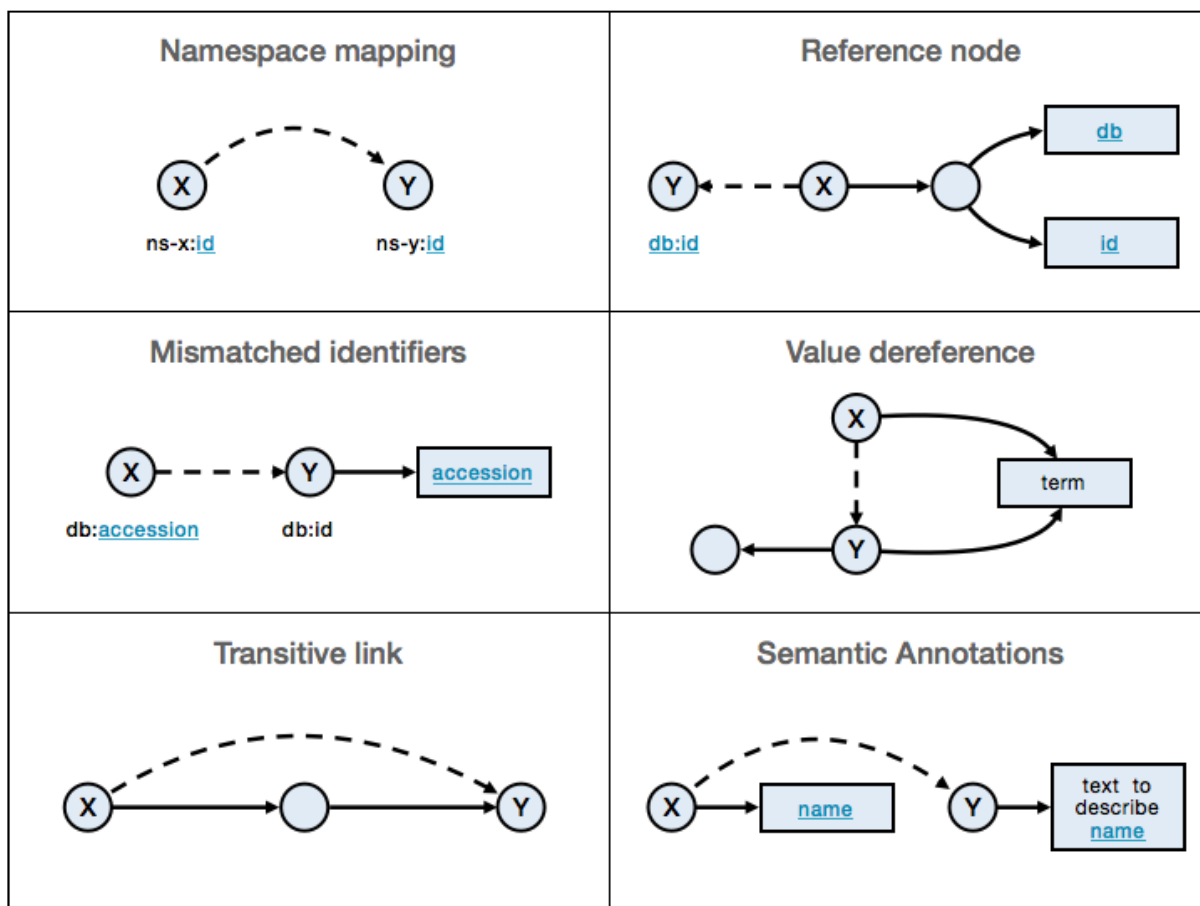


Figure 4 Patterns to align instance level identity over linked data

Table 2 represents all statistics for new explicitly added cross data source mappings. The mappings are now automatically generated as part of the extraction, transformation and loading (ETL) process described in Chapter 3. For the latest information about the explicitly generated cross-data source mappings, please refer to the [1] data sources page.

Source dataset	Destination dataset	Linked Data Mapping Rule	Number of connections	Semantic relationship
BioPax Citations	PubMed	Reference Node (Unification)	1,219,394	skos:exactMatch
BioPax Proteins	UniProt	Reference Node (Unification)	124,706	skos:exactMatch
BioPax Proteins	UniProt	Reference Node (Reference)	37,824	skos:relatedMatch
BioPax Proteins	EntrezGene	Reference Node (Unification)	17	skos:closeMatch
BioPax Proteins	EntrezGene	Reference Node (Reference)	37,824	skos:relatedMatch
DrugBank targets	UniProt	Namespace	4,660	skos:exactMatch



		Mapping		
DrugBank targets (via HGNC references)	EntrezGene (via HGNC references)	Namespace Mapping	1,617	skos:relatedMatch
Freebase concept	UMLS concept	Value Dereference	3,110	skos:closeMatch
MeSH term code	UMLS concept	UMLS code	63,148	lhgdn:umls_code
Pubmed	UniProt citations	Namespace Mapping	922,406	skos:exactMatch
SIDER Side Effects	UMLS concepts	Namespace Mapping	1,685	skos:exactMatch
UMLS concepts	PubMed Abstract and Tiltle	Semantic Annotation	559,096,581	lifeskim:mentions
UMLS concepts (MeSH)	PubMed MeSH terms	Value Dereference	25,813	skos:exactMatch
UMLS concepts (NCBI Taxonomy)	UniProt organisms	Value Dereference	487,085	skos:exactMatch
UMLS concepts (NCBI Taxonomy)	EntrezGene organisms	Mismatched Identifiers	5,891	skos:exactMatch
UniProt GO terms	Gene Ontology term	Mismatched Identifiers	17,778	skos:exactMatch

**Table 3 Cross data source instance mapping statistics**

### 3.5. Relation mappings

The instance mapping stage described in the previous section deals with the proper linking of the RDF resources with respect to the proper semantic inter-linking. Once this stage is completed, the LLD clients should expect that all relevant instances are semantically aligned and efficiently retrievable within a SPARQL query. Still, this feature has a limited applicability since there is no true schema level alignment across the datasets in the warehouse.

The relation mapping stage derives new links between the knowledge base identifiers that describe the functional relation between the two resources. The main objective is to generate directly usable relations for the causality mining interface. Figure 5 illustrates such example visible through the LLD web interface. The *linkedct:criteria* predicate relates clinical trial with literal that encodes inclusion and the exclusion criteria. Unfortunately the extracted disease instances have completely opposite meaning depending their position in the text and if they are listed in exclusion or inclusion section. Thus, a special relation mapping rule is implemented in order to select the correct predicate.



**Trial NCT00103597**  
 Source URL: <http://data.linkedct.org/resource/trials/NCT00103597>

Subject (46) | **Predicate** | Object | All

Download in [JSON](#) | [RDF](#) | [N3/Turtle](#) | [N-Triples](#)

Statements in which the resource exists as a subject. Inference: Explicit and implicit

Predicate	Object
<a href="#">rdf:type</a>	<a href="#">linkedct:trials</a>
<a href="#">rdfs:label</a>	Trial NCT00103597
<a href="#">foaf:page</a>	<a href="http://clinicaltrials.gov/show/NCT00103597">http://clinicaltrials.gov/show/NCT00103597</a>
<a href="#">linkedct:oversight</a>	<a href="http://data.linkedct.org/resource/oversight/49">http://data.linkedct.org/resource/oversight/49</a>

---

<a href="#">linkedct:org_study_id</a>	2004/135
<a href="#">linkedct:lead_sponsor_agency</a>	Royal Brisbane and Women's Hospital
<a href="#">linkedct:criteria</a>	Inclusion Criteria: - Patients residing in Queensland Australia - Age 40-95 - Parkinson's Disease or MSA diagnosed by a neurologist - Symptoms of orthostatic hypotension, as defined by 2 validated questionnaires Exclusion Criteria: - Patients with acute <u>cardiomyopathy</u> or cardiac condition - Patients unable to give consent - Patients not stable on their Parkinsonian medications - Patients with another cause for autonomic neuropathy
<a href="#">linkedct:study_design</a>	Treatment, Randomized, Double-Blind, Uncontrolled, Crossover Assignment, Efficacy Study

Subject	Predicate	Object
NCT00103597	hasExclusionCriteria	<a href="#">umls-concept:C0442874</a>
NCT00103597	hasExclusionCriteria	<a href="#">umls-concept:C0878544</a>

Figure 5 Extracting inclusion/exclusion criteria from between clinical trial and UMLS concepts

Table 4 lists all causality relations currently supported by the LLD service and available in the Causality Mining interface<sup>2</sup>.

Source dataset	Destination dataset	Linked Data Mapping Rule	Number of connections	Semantic relationship
Disease from UMLS	Symptom from UMLS	Causal Relation Extraction	35,505	hasSymptom
Drug from DrugBank	Disease from DISEASOME	Causal Relation Extraction	8,201	treat
Drug from DrugBank	UniProt Protein	Causal Relation Extraction	15,309	hasTarget
Drug from DrugBank or DailyMed	UMLS concepts	Causal Relation Extraction	450,082	hasSideEffect
EntrezGene	UniProt	Causal Relation Extraction	6,033,749	encodeProtein
LinkedCT criteria	UMLS concepts	Semantic Annotation	844,390	hasExclusionCriteria
LinkedCT criteria	UMLS concepts	Semantic Annotation	420,797	hasInclusionCriteria
UniProt and	UMLS concepts	Causal	22,652,496	expressedInOrganism

<sup>2</sup> <http://linkedlifedata.com/refinder>



EntrezGene		Relation Extraction		
UniProt proteins	UniProt keywords	Causal Relation Extraction	275,824	expressedInAnatomicalSystem
UniProt proteins	GO terms	Causal Relation Extraction	5,878,896	hasLocalization
UniProt proteins	GO terms	Causal Relation Extraction	7,885,200	participateInBiologicalProcess
UniProt proteins	UniProt keywords	Causal Relation Extraction	16,912	expressedInCellLine
UniProt proteins	UniProt keywords	Causal Relation Extraction	12,728	expressedInCellType
UniProt proteins	GO terms	Causal Relation Extraction	13,981,106	hasMolecularFunction
UniProt proteins	UniProt proteins	Causal Relation Extraction	68,150,720	binding

**Table 4 Relation mappings generated for the causality mining interface**



## 4. Conclusion

This document presents a second update of the WP7a software prototype. The LLD service shows excellent scalability in terms of data scale and number of supported users. The upgrade of the LLD update process, using data transformers generated by a graphical user interface, lowers the integration effort to add new data sources and allows more efficient automatic updates.

The LLD service fails to directly address the end-users due the limited number of interaction interface. Still, it is one of the best publicly available knowledge bases that aligns large number of biomedical data sources and addresses the structure/semantic heterogeneities by applying instance and schema level alignment.

The LLD prototype is available at <http://linkedlifedata.com> with all the latest software versions and additional resources.



## References

- [1] Linked Life Data Service - <http://linkedlifedata.com>
- [2] Momtchev V. et al., D7a.3.2 Prototype v2
- [3] Miles A., and Bechhover S., SKOS Simple Knowledge Organization System Reference, available at: <http://www.w3.org/TR/skos-reference/>