



## LarKC

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

FP7 – 215535

---

# D7a.1.1

## LarKC Requirements summary and data repository

---

**Bo Andersson (AstraZeneca), Vassil Momtchev (Ontotext)**

Document Identifier:	LarKC/2008/D7a.1.1 /v1.1
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	1.1
Date:	25.09.2008
State:	Final
Distribution:	Public



## EXECUTIVE SUMMARY

The “Requirements summary and data repository” deliverable presents requirements for the semantic data integration in Early Clinical Development use case and the first iteration of the data repository. The use case requirements have been collected in close dialog with the “end users”/scientists at AstraZeneca R&D in Lund.

The objective is to evaluate LarKC as a tool to address the complexity of developing drugs for Chronic Obstructive Lung Disease (COPD). LarKC will be evaluated as a platform for improved integration and interpretation of heterogeneous data, i.e. semantic integration and interpretation of genes-proteins-pathways-target-diseases-drug-patient data.

The case studies from AstraZeneca scientists’ day-to-day work have revealed a need for better tools to integrate and interpret data. Complexities in drug development require tools that support scientist to collaborate and interpret data across scientific domains.

The uncertainties in how to deliver the LarKC platform to scientist require an iterative approach where the user interaction is developed in close collaboration with end users. The LinkedLifeData repository is delivered as an early version in parallel to start the collaborative process of semantic data integration and development of user interfaces.



## 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>	7a.1.1	<b>Title</b>	Requirements summary and data repository
<b>Work Package</b>	<b>Number</b>	7a	<b>Title</b>	Semantic Integration for Early Clinical Development

<b>Date of Delivery</b>	<b>Contractual</b>	M 06	<b>Actual</b>	
<b>Status</b>	version 1.1		final	<input type="checkbox"/>
<b>Nature</b>	prototype <input type="checkbox"/> report <input checked="" type="checkbox"/> dissemination <input type="checkbox"/> other <input type="checkbox"/>			
<b>Dissemination level</b>	public <input checked="" type="checkbox"/> consortium <input type="checkbox"/>			

<b>Authors (Partner)</b>	Bo Andersson (AstraZeneca.); Vassil Momtchev (OntoText)			
<b>Responsible Author</b>	<b>Name</b>	Bo Anderson	<b>E-mail</b>	Bo.H.Andersson@astrazeneca.com
	<b>Partner</b>	AstraZeneca	<b>Phone</b>	












<b>Abstract (for dissemination)</b>	The “Requirements summary and data repository” deliverable presents requirements for the semantic data integration in Early Clinical Development use case and the first iteration of the data repository. The use case requirements have been collected in close dialog with the “end users”/scientists at AstraZeneca R&D in Lund.
<b>Keywords</b>	semantic data integration, early clinical development, pharmaceutical, LarKC, LinkedLifeData, requirements, knowledge management, drug development

<b>Version Log</b>			
<b>Issue Date</b>	<b>Rev. No.</b>	<b>Author</b>	<b>Change</b>
04.07.2008	0.1	Vassil Momtchev	Initial TOC
25.07.2008	0.2	Bo Andersson	Added use case stories (not final)
03.08.2008	0.3	Vassil Momtchev	Revised the user stories
05.08.2008	0.4	Bo Andersson	Added COPD chapter and revised drug development process and user stories
06.08.2008	0.45	Bo Andersson	Revised user stories
12.08.2008	0.5	Bo Andersson	Revised user stories (Scientific feedback)






26.08.2008	0.6	Bo Andersson	Patient safety story updated in collaboration with safety scientists and some cosmetic updates.
12.09.2008	0.7	Vassil Momtchev	Added a uniform structure to all user stories, revised requirement section and added data repository section
15.09.2008	0.8	Bo Andersson and Vassil Momtchev	Refined requirements section, added introduction and conclusion section
22.09.2008	0.9	Bo Andersson and Vassil Momtchev	Review by scientists, corrections and updates added based on their findings.
25.09.2008	1.0	Bo Andersson and Vassil Momtchev	Final changes from review and added executive summary.
25.09.2008	1.1	Bo Andersson and Vassil Momtchev	Updated the deliverable format

## 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@cefril.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 Lab, Sirma Group Corp		Atanas Kiryakov, Ontotext Lab, Sofia, Bulgaria Email: atanas.kiryakov@sirma.bg
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	 Vrije Universiteit	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	 BEIJING UNIVERSITY OF TECHNOLOGY 1958	Prof. Dr. Ning Zhong, THE INTERNATIONAL WIC INSTITUTE, Mabeshi, Japan, Email: zhong@maebashi-it.ac.jp
INTERNATIONAL AGENCY FOR RESEARCH ON CANCER	 International Agency for Research on Cancer Centre International de Recherche sur le Cancer	Dr. Paul Brennan, INTERNATIONAL AGENCY FOR RESEARCH ON CANCER, Lyon, France, Email: brennan@iarc.fr



## TABLE OF CONTENTS

<b>ACRONYMS .....</b>	<b>8</b>
<b>1. INTRODUCTION .....</b>	<b>9</b>
<b>2. DATA INTEGRATION AND INTERPRETATION CHALLENGE IN EARLY CLINICAL DRUG DEVELOPMENT .....</b>	<b>11</b>
2.1. DRUG DEVELOPMENT PROCESS / SEEKING NEW MEDICINES .....	11
2.1.1. TARGET IDENTIFICATION .....	11
2.1.2. HIT IDENTIFICATION .....	12
2.1.3. LEAD OPTIMISATION .....	12
2.1.4. EARLY CLINICAL DEVELOPMENT .....	12
2.1.5. PROOF OF CONCEPT .....	13
2.1.6. DEVELOPMENT FOR LAUNCH .....	13
2.1.7. REGISTRATION AND LAUNCH .....	13
2.1.8. LIFE CYCLE MANAGEMENT .....	14
2.2. THE INTEGRATION AND INTERPRETATION OPPORTUNITY IN EARLY CLINICAL DEVELOPMENT .....	14
<b>3. DISEASE TARGETED FOR EVALUATION - COPD.....</b>	<b>15</b>
<b>4. CASE STUDIES FROM EARLY CLINICAL DRUG DEVELOPMENT.....</b>	<b>16</b>
4.1. IMPROVE THE KNOWLEDGE ABOUT DISEASE AND PATIENTS .....	16
4.1.1. EPIDEMIOLOGIST SYSTEM INTERACTION.....	18
4.2. CLINICAL PROJECT TEAM WORKING ON A NEW TARGET .....	18
4.2.1. CPT TEAM MEMBER SYSTEM INTERACTION.....	19
4.3. IDENTIFYING BIOMARKERS AND TARGET MECHANISMS .....	19
4.3.1. BIOSCIENTISTS SYSTEM INTERACTION .....	21
4.4. SIGNAL EVALUATION OF ADVERSE EVENT REPORTS.....	21
4.4.1. SAFETY EXPERT SYSTEM INTERACTION .....	22
<b>5. REQUIREMENTS .....</b>	<b>22</b>
5.1. METHODOLOGY.....	23
5.2. GENERAL REQUIREMENTS .....	23
5.3. SCALABILITY REQUIREMENTS.....	24
5.4. USER INTERFACE REQUIREMENTS .....	25
<b>6. LINKEDLIFEDATA PROTOTYPE .....</b>	<b>27</b>
<b>7. CONCLUSION .....</b>	<b>29</b>
<b>REFERENCES .....</b>	<b>30</b>



## Acronyms

Acronym	Definition
CD	Candidate Drug
COPD	Chronic obstructive pulmonary disease
GOLD	Global Initiative for Chronic Obstructive Lung Disease (GOLD) [1] works with health care professionals and public health officials around the world to raise awareness of Chronic Obstructive Pulmonary Disease (COPD) and to improve prevention and treatment of this lung disease.
LarKC	Large Knowledge Collider
NDA	New Drug Application
WP	Work Package



## 1. Introduction

The “Requirements summary and data repository” deliverable presents requirements for the semantic data integration in Early Clinical Development use case and the first iteration of the data repository. The use case requirements have been collected in close dialog with the “end users”/scientists. The early prototype of the data repository will be used as a tool to continue the dialog with the scientists/users in an iterative knowledge acquisition process to develop the use case and the LinkedLifeData knowledge base.

Data integration and interpretation is very challenging for pharmaceutical companies in their desire to improve the drug development process and reduce costs [2]. To understand the complexity of diseases and biological processes scientists use new technologies generating huge amounts of data. Tools to fully integrate and interpret all this heterogeneous information do not yet meet their needs. Therefore is semantic data integration and interpretation an opportunity for pharmaceutical companies to increase the productivity by better utilization of all available heterogeneous information.

The “Semantic Integration for Early Clinical Development” use case is about evaluating LarKC. Based on the assumption that it’s no longer possible to use the traditional methods of research where individuals ascertain the information of interest through different file formats, database systems, data semantics and literature [3], the use case objective is to evaluate LarKC as a tool to address the complexity of developing drugs for Chronic Obstructive Lung Disease (COPD). LarKC will be evaluated as a platform for improved integration and interpretation of heterogeneous data, i.e. integration and interpretation of genes-proteins-pathways-target-diseases-drug-patient data.

The aims of our effort are to evaluate if LarKC can:

- Improve the capability to interpret heterogeneous data
- Improve the capability to integrate heterogeneous data
- Improve the capability to capture knowledge
- Stimulate and support new ways of working
- Stimulate and support collaborative working

To measure success we will utilize historical control endpoints where applicable and scientist’s preference, i.e. if user/scientists prefer LarKC to today’s tools we have succeeded.

Section 2 gives you a journey through the drug development process and the data integration challenge. The development process have a long time scale and many different groups are involved in generating data to document and validate the efficacy and safety of a new drug.

In section 3, COPD the disease targeted for the evaluation is introduced. COPD is a severe chronic disease usually caused by smoking. COPD is predicted to be the number three cause of death worldwide by 2020.

In section 4 case studies from AstraZeneca scientist’s day-to-day work are described to explain their need for better tools to integrate and interpret data. These needs will be used to develop the use case for evaluation of LarKC. We will use an iterative development process and continue to involve the scientists/users in the process. (See section on Agile Methods in Deliverable D7b1.1b)

Based on the needs identified, we will in section 5 derive concrete software requirements from the case studies and formalize them in a systematic and consistent way. The requirements are organized in 3 sections:

- General requirements related to the system
- Scalability requirements to discuss the foreseen challenge to integrate huge amounts of triples



- User interface requirements to stress the emerging need for better user interfaces to allow the efficient interaction of non-computer expert users to run advanced queries and analyse large data sets of heterogeneous information

We will also outline our methodology in section 5.

The last section presents LinkedLifeData, the first version to start the collaborative process of semantic data integration. We stress the importance of open and collaborative process in order to achieve the desired level of common understanding and agreement between the different groups. LinkedLifeData initial version is used in two different variants: a public service to host license freely distributable data source and an extended version customized for the use case needs.

## 2. Data integration and interpretation challenge in Early Clinical Drug Development

Every drug is a result of an intensive research and development process performed by teams of scientists driven by a passion to improve the quality of people's lives. This chapter introduces the drug development process and describes the different phases in the process.

The phase in focus for the WP7a use case, early clinical development, is within the pharmaceutical industry recognised as a bottleneck. To understand the complexity of diseases and biological processes scientists use new technologies generating huge amounts of data. Tools to fully integrate and interpret all this heterogeneous information do not yet meet their needs [4]. Semantic integration and tools for interpretation will help scientist grasp relationships between heterogeneous information from diverse sources and collaborate across scientific branches. This can help mitigate the bottleneck in early clinical development including translational medicine<sup>i</sup>.

### 2.1. Drug development process / Seeking new medicines

The drug development process often begins by identifying a new biological target focusing on areas of unmet medical need.

Most pharmaceutical companies use a drug development process similar to the generic one we described in this chapter. The cut between phases and the naming can differ but the content is very similar.

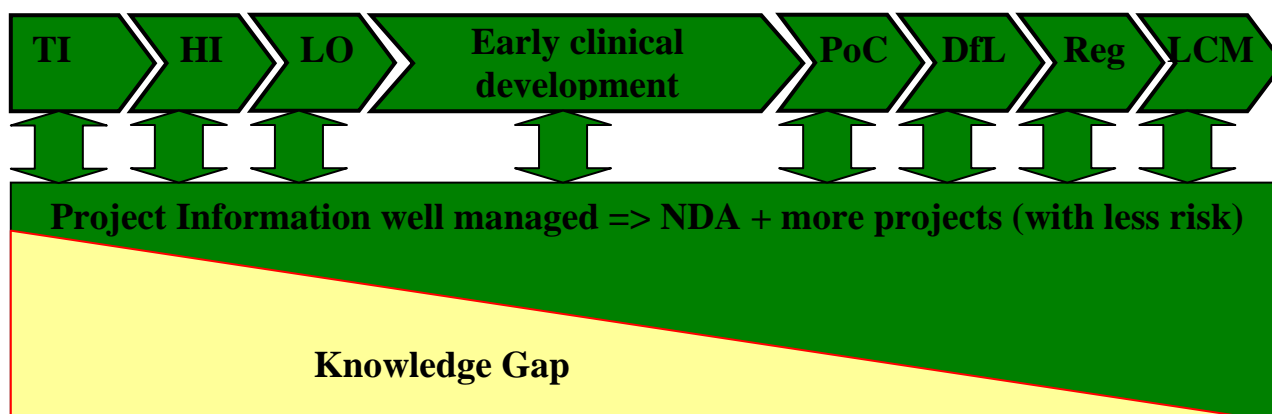


Figure 1, Information perspective of drug development, a journey to fill the knowledge gap (Abbreviations: Target Identification (TI), Hit Identification (HI), Lead Optimisation (LO), Proof of Concept (PoC), Development for Launch (DfL), Registration and Launch (Reg), Life Cycle Management (LCM))

#### 2.1.1. Target Identification

Target identification (and validation) implies a variety of scientific activities to identify biological targets and confirm their potential to be starting points for successful and commercially viable treatments.

Identifying the target engage bioscientists who collaborate with:

- Clinical experts on disease relevance and model development
- Geneticists on genetic relevance and variation

<sup>i</sup> Translational medicine in the meaning of bench to bedside (and the other way around), e.g. prediction for clinical research based on experiments in the laboratory and in silico work.



- Safety experts on concerns related to the target
- Synthetic chemists
- Pharmacology on drugability
- and many more

This broad collaboration required in target identification involves information from a variety of heterogeneous sources. Essential for all coming phases are information about the target and predicted outcome on disease, e.g. mechanism of action and how to measure. Requirements for new biomarkers can be identified.

Case study three about biomarker development will explain how information (decisions) from target identification is of utmost importance in early clinical development.

### **2.1.2. Hit identification**

Next step is to identify “hits” – compounds that are active on the possible targets, which have an element of selectivity and have characteristics that make them suitable to be made into drugs.

The identification is made with help of High Throughput Screening (HTS). High Throughput Screening is an automated system for testing tens or even hundreds of thousands of compounds rapidly and is highly effective for eliminating ineffective compounds and identifying potentially useful ones.

In hit identification heterogeneous information from diverse sources are interpreted to validate hits relevance for the disease pathophysiology, drug-ability from biology and chemical perspective, competitor situation, etc.

### **2.1.3. Lead optimisation**

The identified “hits”, typically in the hundreds, are further refined and their potential as drugs scrutinised. Through hit-to-lead chemistry, hits are converted into a significantly lower number of optimised lead compounds, which are chemicals that have proven to influence the target in a way that gives them the potential to become effective treatments.

These optimised leads are tested for such attributes as absorption, duration of action and delivery to the target, information of utmost importance in early clinical development.

A future opportunity with having information semantically integrated for early clinical development will be for scientists working with hit identification and lead optimisation to access LinkedLifeData to get “feedback” on their work, i.e. being able to see information from later phases.

### **2.1.4. Early clinical development**

To this point scientist have performed a large amount of experiment (using in silico, in vitro and appropriate animal models) to develop high quality compounds. Early clinical development is the phase where discovery meets development where information have to be documented according to Good Laboratory Practice (GLP) and/or Good Clinical Practice (GCP), where responsibility are transferred from discovery to development and where the team are scaled up getting more people involved.

Now it's time to prepare for and test the drug in humans, in early clinical development candidate compounds that have been selected will be further evaluated:

- With regards to safety (Will the drug harm man?)
- Method of drug administration (How will the patient take the product?)
- Bio-availability (How much of the drug will be active in the body?)
- Potential needs for scale-up chemistry (Can we make enough at an acceptable cost?)



A selected compound fulfilling these evaluation criteria is called candidate drug (CD) and will have the potential to be developed into drug.

A CD is further tested in pivotal toxicity studies, in addition to other studies and activities, to provide the necessary regulatory package for submission to authorities to get approval for test in humans.

One example of other studies is scientific measurement studies in diseased populations for biomarker development and evaluation. Biomarkers are very important when developing drugs for slow progressing diseases, e.g. COPD.

Clinical studies, this is testing the drug in humans, can start when approval are received from regulatory authorities. The first clinical study (First Time Into Human (FTIH)) is conducted on a small group of patient or healthy volunteers. This study is the first in a package of phase I studies with the purpose to determine how safe and well tolerated the drug is, if it has the desired effect in human, to learn about pharmacokinetics features of the compound and if the drug has characteristics that would allow it be made into a medicine.

Each activity in early clinical development is very information intensive requiring information from many diverse sources (e.g. access to information from previous phases) and scientists with different expertise. Semantic integration and tools for interpretation will make discovery information available and interpretable for the development project team and facilitate that information from development are available for the discovery team.

#### **2.1.5. Proof of concept**

Now it's time to demonstrate that the concept work in the target patient population. In phase II studies are efficacy, tolerability and safety in the target patient population (100-300) demonstrated and the final dose to be used identified. Further toxicity studies are done to evaluate long-term effects to provide information for the design of long-term clinical studies (phase III).

After this phase everything should be in place to start development for launch, i.e. in the best of worlds the drug should be fully characterized and the concept ready to be confirmed in development for launch.

#### **2.1.6. Development for launch**

In development for launch (phase III) the drug is tested in large populations to confirm its effectiveness, tolerability and safety, monitor side effects and compare it to existing treatments.

Large-scale clinical phase III studies, with anywhere from 500 to more than 10,000 patients, is performed where the drug is tested for rare side effects, effectiveness and safety, and if the medicine is more or less effective in various patient segments (demographic groups).

Information gathered in development for launch is an essential component in the new drug application (NDA). However, this information is also important input for previous phases, e.g. early research in target identification. Semantic integration and tools for interpretation would facilitate good feedback.

#### **2.1.7. Registration and launch**

For registration, results from pre-clinical and clinical studies, the quality data and description of manufacturing process are compiled into a NDA and submitted for review of the regulatory authorities.

The NDA is an extensive collection of information compiled and customized for the authorities.



If approved, the drug can be made commercially available.

### **2.1.8. Life cycle management**

When the drug is on the market life cycle management is important. As through the whole process, safety is in focus. Safety surveillance is done through constantly monitoring adverse effects reported in relation to the drug.

In addition clinical Phase IV studies are performed to examine medical long-term effects, health economic aspects, new indications and possible new formulations of the substance.

Information is generated over the drugs whole lifetime. Semantic integration and tools for interpretation have a potential to help also in life cycle management.

### **2.2. The integration and interpretation opportunity in early clinical development**

The drug development process is about developing a compound and the knowledge to prove its behaviour in a target patient population. Our focus is early clinical development where diverse scientists meet to predict behaviour in human based on “existing” knowledge and knowledge/information from experiments using in silico, in vitro and appropriate animal models.

The breadth and depth of relationships between heterogeneous information a drug project need to conceptualise in early clinical development is almost impossible to handle with present tools [5].

### 3. Disease Targeted for Evaluation - COPD

COPD [6] is a severe chronic disease usually caused by smoking. COPD is predicted to be the number three cause of death worldwide by 2020 by the Global Burden of Disease Study.

The definition of COPD adopted by GOLD [1]:

*A disease state characterized by airflow limitation that is not fully reversible. The airflow limitation is usually progressive and associated with an abnormal inflammatory response of the lungs to noxious particles and gases.*

COPD is complex disorder that to a large extent comprises:

- **Chronic bronchitis**, whose clinical definition is productive cough (from bronchial secretion) on most days for 3 months/year for 2 consecutive years. The mucus hypersecretion comes from hypertrophied bronchial glands and increases the risk of bacterial lung infections.
- **Emphysema**, which has a pathological definition with enlargement of the alveoli due to the destruction of the walls between them. These walls contain elastic fibres, so their destruction reduces the elasticity of the lung, leading to collapse, and thus obstruction, of airways.

Early diagnosis and effective treatment of COPD remains unsatisfactory and an unmet medical need. Central to meet the requirement is better understanding of the pathophysiological changes that occur over time in response to the long-term use of cigarettes. The clinical manifestations of COPD are only recognized by the patients at the later stages of disease when loss of lung function begins to deteriorate their quality of life.

These symptoms include a broad range of biological insults including chronic bronchitis, emphysema, dyspnoea, systemic inflammation, cardiovascular deterioration, loss of muscle mass and wasting, and susceptibility to respiratory infections causing exacerbations.

COPD has been recognized as a syndrome rather than a disease, with important extra-pulmonary manifestations that impact on many different organs. These co-morbidities<sup>i</sup> have major effects on health outcomes, e.g., cardiovascular disease and lung cancer are the two leading causes of morbidity and mortality in COPD patients [8]. The appropriate assessment and management of these systemic complications of COPD are critically important for the improvement in future treatment and outcome of COPD. The combination of complexity, severe co-morbidities and unmet medical needs for early diagnosis and treatment in COPD require new approaches to be solved. LarKC will improve information access and interpretation with help of semantic integration and computational support to conceptualise this complexity. Case study one focuses at the need for better understanding of COPD.



**Figure 2, COPD patient** (from European Lung Foundation<sup>ii</sup>)

<sup>i</sup> Co-morbidity, more than one medical condition (disorder) existing simultaneously in a patient.

<sup>ii</sup> <http://www.european-lung-foundation.org/index.php?id=27>



## 4. Case studies from early clinical drug development

This chapter outlines the use case requirements in multiple case studies. Every story specifies the user goals, the steps to interact with the system and the anticipated results. The layout of the stories is first to introduce the high-level user needs by shortly overview the typical tasks (this may includes from daily tasks like I have to map the results of program X to program Z, to few month project as the preparation of report over the link between two specific diseases). The case studies specify the set of system features seen from user perspective. The interactions with the system are described as non-formal textual stories.

AstraZeneca R&D Lund is one of the leading centres in the world for development of respiratory drugs and treatment. Hence, all case studies evaluate the system with respect to COPD and other related diseases. To succeed in developing innovative treatment for COPD, filling the gap of knowledge in understanding the etiology and pathophysiology of COPD and how the human body works is crucial. Scientists with a variety of different expertises, e.g. biology, genetics, chemistry, pharmacokinetics, epidemiology and medicine, work in collaboration to develop new innovative drugs. The work includes a need to develop the disease understanding and knowledge about patient segmentation. LarKC can provide a common knowledge repository for the drug development process, the LinkedLifeData and computerized tools for interpretation. This has the potential to improve the success rate in early clinical development by improving our capability to integrate and interpret information from diverse sources and facilitate new ways of working across scientific branches.

The focus of the case studies is on COPD and related diseases but can easily be generalized to other disease areas.

The case studies will evolve during the evaluation process. Each story tries to capture our current understanding of the user needs. It is very likely that the user needs may change after start utilising any piece of software (see section on Agile Methods in Deliverable D7b1.1b).

### 4.1. *Improve the knowledge about disease and patients*

Improving the understanding of the etiology and pathophysiology of COPD and patients different responses [7] to treatment is a continuous task. Scientists employ a range of observational to experimental methods to develop this knowledge.

Epidemiologists have a central role in the work to improve disease and patient understanding. Their expertise of studying various factors affecting health in large populations, real life, is vital. The long-term objective is to develop enough knowledge to be able to apply a long-term holistic view on COPD where prevention, early diagnosis, co-morbidity with cardiovascular diseases and lung cancer are all brought into the big picture [8].

In this story we overview the continuous knowledge building process of drug development, where the researcher has to systematize heterogeneous information and experimental results into knowledge about a disease and its patient.

**Problem description:** Identifying causal relationships between environmental factors, genetic heritage, biology in disease processes, patient groups and outcomes are important aspects of our COPD understanding. Developing knowledge about these relationships requires integration and interpretation of heterogeneous information and relationships from a vast number of diverse data sources. The structure of the data sources may vary from database entries to annotations derived from unstructured text. Current tools are not developed to handle the amount of diverse data sources required. The epidemiologists require an

environment that is better suitable for a continuous improvement of the disease knowledge. The main questions in focus are:

- Explore biological and environmental risk factors for developing a disease and prognosis for the patients with it
- Explore hypotheses about common mechanisms that relate multiple diseases (explain co-morbidity)
- Explore the hypotheses for casual chains of a disease (early diagnosis)
- Explore hypotheses about patient characteristics and environmental factors that helps characterize clinically relevant patient groups

**Example system interaction:** The epidemiologist will test hypotheses (develop knowledge) relevant for patient segmentation and the causality chains behind development of COPD and related diseases. The workflow is iterative (continuous) and new data sources will be integrated when required. The outcome will continuously be added back to the knowledgebase within respiratory. The disease knowledgebase will then influence strategies and be available as a tool in the day to day work by project teams and individual scientists. Topics of interest are hypotheses about common mechanisms that relate multiple diseases (explain co-morbidity):

- Biological and environmental risk factors for developing a disease and prognosis for the patients
- Hypotheses for casual chains of a disease (early diagnosis)
- Hypotheses about patient characteristics and environmental factors that can explain segmentation criteria

**Success criteria:** Reduce the time to test new hypothesis for patient segmentation and the causality chains behind the development of related diseases. We will utilize historic control endpoints for evaluation, however scientists' preference and subjective judgment will be the final measure of success.

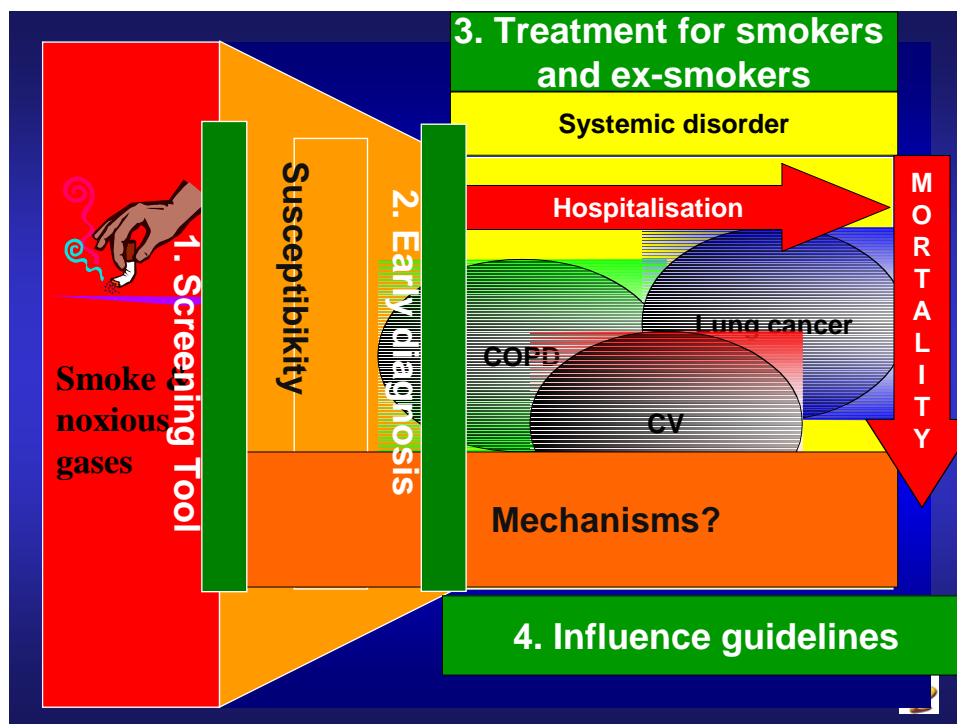


Figure 3 New holistic treatment regime including prevention and chronic condition management



#### 4.1.1. Epidemiologist system interaction

The section describes the conceptual understanding of an epidemiologist for functionality LarKC platform has to realize. The proposed workflow is subject of discussion in the requirements section and may not be fully consistent with respect to software engineering approaches.

Plugin	Description
<b>Retrieval</b>	Retrieve structured or semi-structured data sources and transform them into RDF if distributed in other formats.
<b>Abstraction</b>	Apply algorithms for URI mapping, where different naming conventions are used.
<b>Select</b>	Select subset or full knowledge base based to operate with.
<b>Infer</b>	Remove redundancy with relation interpretation (e.g., use database cross reference to identify equivalent concepts; interpret semantic networks to map different identifiers).
<b>Decide</b>	Navigate the structured information and explore the different relations and their sources.
<b>Information extraction pipeline that operates over the ontology instances collected by the previous process.</b>	
<b>Retrieval</b>	Retrieve unstructured document and transform the document content (e.g., Medline abstract and content is encoded as literal) and meta-data RDF.
<b>Abstraction</b>	Apply named-entity recognition, semantic annotation to the ontology instances and relation extraction.
<b>Select</b>	Select subset or full knowledge base based to operate with.
<b>Infer</b>	Perform consistency checking of the extracted relations against the knowledge from the structured data sources.
<b>Decide</b>	Specify queries over to be evaluated against the structured and unstructured information, explore the different entities and the co-occurrence among them.  Extend the existing knowledge base (e.g., with textual document) and improve the disease understanding.

Table 1 Epidemiologist system interactions.

#### 4.2. Clinical project team working on a new target

The early Clinical Project Team (CPT) has the challenging responsibility to bring new candidate drugs through to proof of concept. They start working with the new compound together with the discovery team in the lead optimisation phase. The CPT has to overcome the complexity:

- Of planning the development program
- Developing the regulatory package
- Safely test the drug first time in human
- Developing new methods and biomarkers (e.g. methodology studies using imaging and tissue acquisition)
- To prove the drug principle

These tasks require continuous interactions with different experts and integrated access to many heterogeneous information sources. For more information please refer to chapter 2.1.4 Early clinical development.

**Problem description:** The CPT has the delicate challenge to plan the work, execute studies and respond with accuracy to findings and external questions. They need to be on top of all available information to



handle the uncertainty that exists in Early Clinical Development without jeopardizing neither the patients' safety nor the project. The clinical team have to efficiently access knowledge about:

1. The disease and patient segmentation
2. Risk factors for drug class and/or biological targets
3. How other companies or institutes design their studies and programs
4. Patient availability
5. How animal models will scale to human
6. Previously performed studies internally and externally with characteristics of value for this program
7. Failed experiments

There is no available solution to handle all these heterogeneous information sources efficient for the project. Therefore are the information integration and interpretation a major obstacle for the CPT, as they have to grasp knowledge from heterogeneous information sources and work with many groups.

**Example system interaction:** The CPT team respond to a question about the drug by investigating all related clinical studies by specifying different search criteria like dose/drug concentration, adverse reactions, inclusion/exclusion patient criteria and etc. The search has to cover all public and in-house information sources and resolve the differences in the used terminology.

**Success criteria:** Improve confidence and reduce time when responding to the unexpected questions and decrease the number of studies and/or subjects. Another benefit is the introduction of "project memory" to facilitate the navigation of the full project history and being able to easily provide authorities with summaries e.g. on a new target in educational purpose. The CPT's preference to use LarkKC will be the final measure.

#### **4.2.1. CPT team member system interaction**

Again this section presents the conceptual understanding of the CPT about the system and the needs it has to address. It should be taken as recommendation about the required functionality rather used as design guideline. CPT team members focus over the completeness of the information and need to be sure that they do not omit existing clinical trials and the need efficiently retrieve them using the applied doses and measurements.

<b>Plugin</b>	<b>Description</b>
<b>Retrieval</b>	Reuse the knowledge generated in case study one and extend it with the results from the epidemiologist reports.
<b>Abstraction</b>	Apply information extraction algorithms to try identifying the mechanisms of already conducted clinical trials (e.g., extract the medicament doses and the target groups).
<b>Select</b>	Filter the available sources.
<b>Infer</b>	Use faceted search interface to reduce the number of document by applying multiple resource classification (e.g. documents to discuss respiratory diseases).
<b>Decide</b>	Interact with user friendly interface (natural language queries may be an option) to display the results.

**Table 2 CPT member system interactions.**

#### **4.3. Identifying biomarkers and target mechanisms**

Bioscience contributes in several phases of the drug development process, e.g. Target Identification and Early Clinical Development. Bioscientists uses diverse technologies, e.g. both protein (proteomics) and



transcript expression analyses, that generate huge amounts of experimental data in differing electronic formats that describe both qualitative and quantitative measurements, e.g. of proteins and images (tissue). This work involves laboratory analyses of samples taken from individual patient under specific condition, interpretation of the results in the context of disease pathophysiology, patient segments and the development of *biomarkers* [9]. One of the common goals today in Bioscience projects is to define certain biomarker standards that can be used to:

- Classify or group patients within a disease spectrum – disease evolution
- Identify responders
- Identify safety signals
- Predict outcome

*Biomarkers, a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention* [9], is an important “tool” for early clinical development and later phases of drug development to be used e.g. as a diagnostic tool for the identification of a unique patient segment, for staging or classification of disease, as an indicator of disease prognosis, prediction and monitoring of clinical response to an intervention.

**Problem description:** Huge amounts of experimental data are generated in the different electronic formats that describe both qualitative and quantitative measurements, e.g. proteins and images. Data interpretation is a non-trivial process that requires overcoming:

- Syntax differences in the generated result formats
- Semantic difference in the formats, e.g. used identifiers
- Verify, validate and compare the experimental results with other established data sets
- Vast heterogeneity of the interpreted information
- Efficient secondary usage of past experimental results and analysis conducted in the later phases of the drug development research

Available tools do not fulfil the Bioscientists needs of multi-aspect analysis capability and integration of large amounts of heterogeneous information, i.e. there is a problem in the interpretations of data.

**Example system interaction:** The Bioscientists needs efficient knowledge system to test various hypotheses for further experimental work. For instance such hypothesis could be “COPD patients express into their plasma special proteins that are the result of destructive processes associated with smoking”.

1. Generate plasma sample from COPD patient cohorts and experimentally generate a protein profile composed by thousands of separated individual proteins
2. Identify the individual proteins and map them to public data sets to describe protein sequences
3. Analyse the new relationships with respect protein identity, metabolic pathways, molecular interactions, signalling pathways, gene regulation and genetic interactions
  - a. Answer good enough? If (NO) add information/annotations and return to (1)
4. Perform experiment (Re-analyse plasma samples using alternative methods of separation, fractionation) and upload the experimental data set results
5. Verify if the experimental result support the hypotheses. Validation by using another cohort of samples collected from other COPD subjects. *If requirement not fulfilled - start new iteration*

**Success criteria:** Narrow down the uncertainty and reduce the number of needed experiments, by comparing the different studies. Decrease the time to analyse the experimental results. Historic control endpoints will be used to evaluate success in parallel with user preferences.



### 4.3.1. *Bioscientists system interaction*

The bioscientists’ workflow recommendation used the same sequence of events as the previous two stories. It stresses over the requirements for easy interaction with the system based on the results obtained from the laboratory experiments.

Plugin	Description
Retrieval	Import laboratory output results in the form of excel or CSV file. Apply simple algorithms for transformation to RDF.
Abstraction	Identify the proteins and map them to public data sets (e.g., create relation to protein sequence resources).
Select	Filter the available sources or remove some of the results.
Infer	Apply algorithms to interpret the semantics of the relation and annotate the results.
Decide	Test hypotheses based on the imported results.

**Table 3 Bioscientist system interaction.**

### 4.4. *Signal evaluation of adverse event reports*

Patient Safety is responsible for the safety of our drugs. A safety expert manages adverse event reports and defines the safety profile of our drugs and targets. Part of that is to define the underlying COPD patient profile from a safety perspective, which links to epidemiology – co-morbidity, concomitant medication, adverse event profile in placebo patients and during use of current drugs etc. (see the first case study).

An adverse event is any untoward medical occurrence in a patient or subject administered a drug and which does not necessarily have a causal relationship with this treatment. Adverse events are collected both in clinical trials and post marketing of a drug. When the safety experts receive an adverse drug event reports they need to evaluate the medical occurrence and verify if there is a causal relationship between the drug and the event. This is done both for individual reports and for aggregated reports. Thus, the process requires the expert to integrate and to interpret massive amounts of heterogeneous information. The process of pharmacovigilance and patient risk management consists of the following steps:

1. Signal detection, usually based on adverse event reports
2. Signal evaluation – causal relationship between drug and event, or not
3. Risk mitigation if causal relationship – what can be done to minimize risk
4. Assess effect of risk mitigation activities – simulation based on previous experience or direct measurement after mitigation activities have been implemented

**Problem description:** Signal evaluation is a critical step in pharmacovigilance. To evaluate the medical occurrence and do a causality assessment of the event require heterogeneous information to be integrated and interpreted.

**Example system interaction:** During signal evaluation, the safety expert will evaluate the medical occurrence and verify if there is a causal relationship between the drug and the adverse event. The evaluation includes screening for other causes than the drug itself, such as the role of concomitant therapies and non-drug-related causes. One method for drug causality assessment is called the RUCAM [10] and consists of 7 criteria:

1. Time to onset of the reaction
2. Course of the reaction
3. Risk factor(s) for drug reaction



4. Concomitant drug(s)
5. Non-drug related causes of event
6. Previous information on the drug
7. Response to readministration

Four of the RUCAM criteria (number 3-6 above) are considered especially suitable to evaluate through searching and integrating large amounts of heterogeneous data.

**Success criteria:** Improves quality and/or speed when evaluating safety signals. We will utilize historic control endpoints for evaluation (if applicable), however scientists' preference and subjective judgment will be the final measure of success.

#### 4.4.1. Safety expert system interaction

The safety expert is the last system interaction, but definitely not the least important one. The safety is a top-priority for all pharmaceutical companies. A similar workflow to the other stories could be applied to the LarKC system understanding from the user perspective.

Plugin	Description
<b>Retrieval</b>	Collect all resources to cover the safety of a drug including public safety regulatory reports and previously clinical studies or alerts.
<b>Abstraction</b>	Perform named-entity recognition of the available adverse events.
<b>Select</b>	Filter the selected data sources and/or find the most relevant one.
<b>Infer</b>	Infer the relation of the different adverse events described in the coding systems.
<b>Decide</b>	Use a friendly interface to present the result and explore the full information sources generated from the previous three stories.

Table 4 Safety expert system interaction.

## 5. Requirements

The chapter formalizes a list of requirements towards the Early Clinical Development use case prototype. All case studies demonstrate the urgent need of computational tools for semantic data integration and interpretation to help scientific experts to interpret heterogeneous information sources collectively. They need tools and systematized approach to integrate and interpret information and linking information for genes-proteins-pathways-target-diseases-drug-patient. We have identified a couple of important needs that the use case prototype and LarKC platform have to address:

- A highly scalable knowledge base where structured heterogeneous information used during the drug development process is incrementally integrated
- Linkage of unstructured textual documents and annotations to the knowledge base resources using information extraction algorithms
- Advanced semantic search/navigation capability
- Computerized support to interpret information
- Computerized support capturing learning (annotating) and adding to data repository
- Continuous access to the updated knowledge repository when they search for relations in articles/documents



We also expect LarKC to stimulate teams to find new ways to work, e.g. where the present sequential process is not optimal use a more iterative process based on the opportunities provided through more accurate access and interpretation of information and knowledge. This section specifies the initial list of formal requirements towards Early Clinical Development use case prototype to be realized with the LarKC platform.

All functional and non-functional requirements towards the prototype and LarKC platform are organized in three categories.

### **5.1. Methodology**

The case studies were developed in close collaboration with AstraZeneca scientists in Lund. We started with a series of workshops and have continued to interact with them to finalise the first iteration of case stories.

The requirement for better support to integrate and interpret heterogeneous information is unanimous. However the experience how to best utilize semantic integration and interpretation are limited, both among the scientists and in computer science. Therefore we will work very close with scientists representing the four case studies in an iterative process to develop the use case requirements and the LinkedLifeData prototype. An important component in this process is access to the LinkedLifeData prototype to be able to capture user experience.

The main objective with the use case is to confirm that the LarKC platform fulfil its requirement and develop value to Early Clinical Development through semantic data integration and interpretation. To measure how well the objective is fulfilled each case study will develop measures. The main criteria will be scientists' preference, i.e. if the scientists choose LarKC as preferred tool in the case stories we have succeeded. We will also use historic control endpoints where applicable.

### **5.2. General requirements**

This section presents the initial list of requirements for the first prototype. LarKC platform must realize a software infrastructure to build a scalable knowledge base to support semantic integration and efficient interpretation of heterogeneous information type.



**[R 1] Efficient RDF data model support**

The early clinical drug development process is a very information intensive process involving the consolidations of knowledge from many diverse sources. RDF data model is the only feasible choice for integration of heterogeneous semi-structured cross-domain knowledge, redundant domain knowledge and semantic annotations of unstructured document. For scalability estimation of the please refer to 5.3.

**[R 2] Information provenance, different contexts and custom meta-data**

Data quality of the different data source is critical aspect in the correct results interpretation. The knowledge base must support logical fragmentation of the information based on the different data sources. It must be possible to go to the source that originates specifics statement and associate additional meta-data like annotation creator or timestamp.

**[R 3] Import and export of data sources between different knowledge bases**

There are multiple cases where different version of the knowledge base must be maintained. For instance it must be possible to replicate data sources from the public knowledge bases to privately owned systems.

**[R 4] Apply information extraction algorithms and generation of semantic annotations over unstructured textual documents**

Ontology annotations are still not widespread in the biomedical domain. Most of annotations are still published as free text. Large number of textual data sources partially or completely lack semantic annotations to support the identification of specific terms in the text. Named-entity recognition and instance unification is the only possibility to support interoperability of large number of biomedical data sets.

**[R 5] Identity aggregation**

The different databases introduce local identifiers that often refer one and the same entity. Also, there is no universal accepted convention for biological entity naming (i.e., LSID vs. PURL), so the prototype needs computational support to efficiently aggregate redundant identifiers.

**[R 6] Robust backup and restore procedures with acceptable resilience**

A simple mechanism for knowledge base snapshot backup must be supported. The backup must be able to restore the knowledge base to a previous snapshot state, where all data queries return repeatable results.

**[R 7] Security model**

The prototype must provide security mechanisms to authenticate and log the activity of the system users. The authenticated users will have access to full knowledge base. A separate role for administrator must also be available.

### **5.3. Scalability requirements**

The scale of the knowledge base in the sense of the number asserted facts is depended to the type of expected questions and completeness of the anticipated answer. It is very difficult to set direct and objective measurement parameters to estimate the results before the first evaluation phase. Hence, in this section we present direct or indirect factors and results statistics derived from the initial prototype to measure the minimum of the scale of the system to pose significant interest to drug development researchers.

The two major sources of triples are the transformed structured databases and the semantic annotation of unstructured documents. The number of custom annotations is expected to be relatively small compared to the overall number of triples. In the following paragraphs we set initial boundaries for the expected



knowledge to be integrated in the use cases. Based on the input of the researchers during first prototype evaluation testing the need of new information sources is likely to increase.

Release 14.1 of Swissprot (the curated part of Uniprot database) from 2 September 2008 has 397,539 entries [11] or 320 million RDF statements. We can expect a steady growth in the number of annotations like that for the last 3 years to continue and keep the 50% increase. Thus, the scale of curated protein sequence information in 3 years is estimated to be at least 480 million statements. A similar information scale is expected for the gene-specific databases like Entrez-Gene and Ensembl. The pathways databases measured with the scale of files of BioPAX [12] are less than 100 million statements, but are likely to increase with the extended support with more databases. The scale of structured information for patient, drug and disease information will be significantly lower compared with previous categories, since most of the existing information is distributed in form of semi-structured or unstructured text and it will be calculated as semantic annotations.

**[R 8] Support of knowledge base and data retrieval of at least 1 billion explicit statements generated from structured data sources**

This requirement presents the minimal scale of the repository to be used for evaluation. The size of unstructured data sources like Medline publication, clinical trials and internal document could vary significantly depending on the new emerging requirements. Our initial tests have concluded that for 1.2 million random Medline abstracts. According to Medline statistics for 2007, 670'000 new articles have been added [13], so to give estimation for the scale this is the full size of articles to be added only for two years. Gazetteer initialised with SNOMED controlled vocabulary plus drug substance concepts and ABNER [14] generated 10,884,032 annotations of type diseases/disorder, drug, gene, proteins, DNA, RNA, cell line and type. For full text articles we can expect that proposed number will increase at least by factor 20. We can estimate that at least 5 statements would be required to describe the position and type and the possible ontology instance of the individual annotations.

**[R 9] Minimum support of 1 billion statements annotation generated as result of information extraction process**

We expect the scalability requirements to constantly **increase** with the evolution of the knowledge base for Early Clinical Development.

#### **5.4. User interface requirements**

**[R 10] User interaction that stimulate contribution**

Support the user to store results back into the knowledge base in a structured format.

**[R 11] Flexible user interaction.**

Different users and different tasks require different interfaces. This can result in more than one interface. Iterative development processes have to be applied, we don't know how to best utilize the semantic to help/stimulate users to interpret/navigate in the best possible way.

**[R 12] Export of result**

The scientists need to be able to export the outcome from a LarKC interaction. Exactly how must be evaluated with help of the prototype.

**[R 13] High level query language**

Users need to be able to use high-level query language, e.g. Natural Language.



**[R 14] Administrative console to control the different data sources**

An administrative console to monitor the different data source, the active user processes and the overall system state is needed for proper system administration. The console must provide possibility for the insertion of new data sources and deletion of existing ones.

**[R 15] Mechanisms to import user structured data sets in formats like Excel or tab delimited format and transform to RDF**

The system user interface must provide a convenient way to import custom data sets to the system and provide mechanisms for their transformation to RDF.



## 6. LinkedLifeData prototype

This chapter presents the early prototype of semantic integration knowledge base to be applied for Early Clinical Development. Chapter 2.1 presented the drug development process as a very intensive knowledge exchange process where the individual projects takes 10 or more years, a significant period, enough to expect even shift in the computing practices and paradigms. LinkedLifeData is a term coined to describe the developed methodology in accordance of the principles of Linked Data initiatives and the software to implement data repository itself, but not the actual data and schema representation. We foresee that LinkedLifeData could be applied to different life science use cases, where universal ontological representation will be difficult to get broad community agreement. Thus, multiple variants to capture the specificity of data sources and possibility for semantic integration may be supported.

Objectives for the initial prototype are to validate the feasibility of LinkedLifeData approach and improve our understanding about the researchers need and the applied integration methodology. The current variant is labelled LinkedLifeData – Pathway & Interaction Knowledge base (PIKB), which map of different sources is based on common reference identifiers, powered by rule-based inference strategy to provide additional implicit knowledge about initial concepts of integrated sources. On a high-level LinkedLifeData – PIBK is foreseen to be integral part of LinkedLifeData – Translation Medicine, the ultimate goal of our work.

### Figure 4 LinkedLifeData current road map

The current version of the data repository is in an initial stage. We expect major changes in the existing schema and data sources during the PIBK prototype development (M6-M18). The gathered evaluation results and requirements analysis will be used as input for the next iteration. In the list below we present the integrated data sources, used dataset (e.g., full database or parts of it), applied schema and a short overview of the included type of resources.



Database	Dataset	Schema	Description
<b>Uniprot</b>	Curated entries	Original by the provider	Protein sequences and annotations
<b>Entrez-Gene</b>	Complete	Custom RDF schema	Genes and annotation
<b>iProClass</b>	Complete	Custom RDF schema	Protein cross-references
<b>Gene Ontology</b>	Complete	Schema by the provider	Gene and gene product annotation thesaurus
<b>BioGRID</b>	Complete	BioPAX 2.0 (custom generated)	Protein interactions extracted from the literature
<b>National Cancer Institute - Pathway Interaction Database</b>	Complete	BioPAX 2.0 (original by the provider)	Human pathway interaction database
<b>The Cancer Cell Map</b>	Complete	BioPAX 2.0 (original by the provider)	Cancer pathways database
<b>Reactome</b>	Complete	BioPAX 2.0 (original by the provider)	Human pathways and interactions
<b>BioCarta</b>	Complete	BioPAX 2.0 (original by the provider)	Pathway database
<b>KEGG</b>	Complete	BioPAX 1.0 (original by the provider)	Metabolic pathways
<b>BioCyc</b>	Complete	BioPAX 1.0 (original by the provider)	Metabolic pathways
<b>NCBI Taxonomy</b>	Complete	Custom RDF schema	Organisms

**Table 5 Initial list LinkedLifeData prototype data sources**

A public demonstration service is available at:

<http://www.linkedlifedata.com>.



## 7. Conclusion

The case studies from AstraZeneca scientists' day-to-day work have revealed a desire for better tools to integrate and interpret data. Complexities in drug development require tools that support scientist to collaborate and interpret data across scientific domains.

Requirements extracted from the case studies are feasible to meet with the LarKC platform. We will evaluate if LarKC's capability to semantically integrate and reason over huge heterogeneous and incomplete data can fulfil scientist's requirements and help to solve the COPD mystery. This will be measured by historic category endpoints and user preferences. The diversity of data needed to be integrated and interpreted spans genes-proteins-pathways-target-diseases-drug-patient data.

Scientists desire to evaluate if LarKC can:

- Improve the capability to interpret heterogeneous data
- Improve the capability to integrate heterogeneous data
- Improve the capability to capture knowledge
- Stimulate and support new ways of working
- Stimulate and support collaborative working

In other words, the use case "Semantic data integration in Early Clinical Development" will be used to evaluate the LarKC platform. The scalability challenge raised by this use case will be a major test for the approach and will be used as fair evaluation for its applicability in pharmaceutical industry. The success criteria are that LarKC can help scientists to integrate and interpret huge heterogeneous and incomplete data. This can be an interesting win-win opportunity.

The uncertainties in how to deliver the LarKC platform to scientist require an iterative approach where the user interaction is developed in close collaboration with end users. The LinkedLifeData repository is delivered as an early version in parallel to start the collaborative process of semantic data integration and development of user interfaces.



## REFERENCES

- [1] From the Global Strategy for the Diagnosis, Management and Prevention of COPD. Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2007. Available from: <http://www.goldcopd.org/>
- [2] Sokoll, K. Optimizing Drug Development Strategies.  
<http://www.pharmabioingredients.com/articles/2006/11/optimizing-drug-development-strategies>.
- [3] Stephens, S. Effective data integration is of increasing importance within the life sciences. 2007.  
<http://www.semantic-web.at/10.36.162.article.dr-susie-stephens-effective-data-integration-is-of-increasing-importance-within-the-life-s.htm>.
- [4] FDA white paper Innovation or Stagnation (March 2004): “developers have no choice but to use the tools of the last century to assess this century's candidate solutions.” “Industry scientists often lack cross-cutting information about an entire product area or information about techniques that may be used in areas other than theirs”
- [5] Muggleton. 409, s.l. : Nature, 23 March 2006, Vol. 440.
- [6] <http://en.wikipedia.org/wiki/COPD>.
- [7] Connor, Steve. Glaxo Chief: Our Drugs Do Not Work on Most Patients.: Independent/UK, December 8, 2003.
- [8] Gerhardsson de Verdier, M. The Big Three Concept - A Way to Tackle the Health Care Crisis? Proc Am Thorac Soc Vol 5. pp 800–805, 2008
- [9] Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Group, Definitions Working: Clin Pharmacol Ther, Vol. 69,  
<http://www.nature.com/clpt/journal/v69/n3/pdf/clpt200113a.pdf>.
- [10] Causality assessment of adverse reactions to drugs--I. A novel method based on the conclusions of international consensus meetings: Application to drug-induced liver injuries. Gaby Danan, Christian Benichou. 11, s.l. : Journal of Clinical Epidemiology, November 1993, Vol. 46.
- [11] UniProtKB/Swiss-Prot protein knowledgebase release 56.1 statistics.  
<http://expasy.org/sprot/relnotes/relstat.html>.
- [12] BioPAX : Biological Pathways Exchange . <http://www.biopax.org/>.
- [13] Medline Fact Sheet. <http://www.nlm.nih.gov/pubs/factsheets/medline.html>.
- [14] ABNER: A Biomedical Named Entity Recognizer.