LarKC: Semantic Data Integration for Early Clinical Development

Vassil Momtchev (Ontotext)

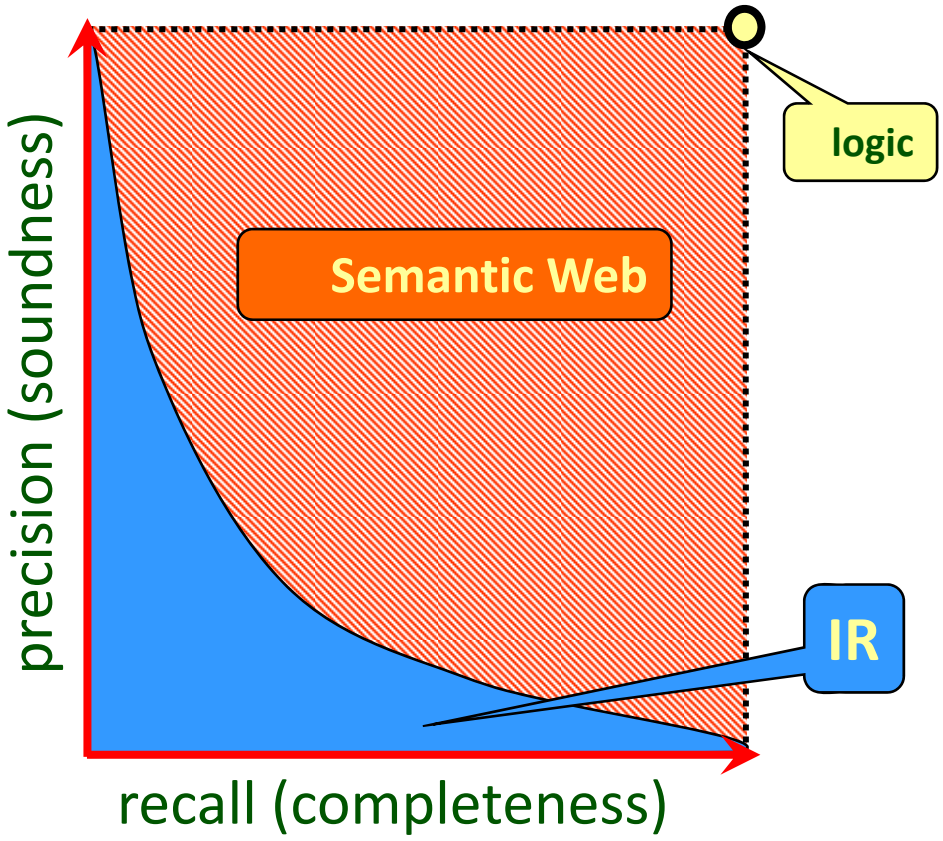
Bosse Andersson (AstraZeneca)

Large Knowledge Collider in a Nutshell

- LarKC goes beyond the current limited storage, querying and inference technology
- Fuse reasoning with search aims for a paradigm shift
- Achieve “Web Scale Reasoning”

Web Scale Reasoning

- "Web Scale and Style Reasoning"
  - Giving up 100% correctness:
  - trading quality for size
  - often completeness is not needed
  - sometimes even soundness is not needed



Main Innovations

- Enriching current logic-based Semantic Web reasoning
- Employing cognitively inspired approaches and techniques
- Achieve scalability trough giving up completeness
- Achieve scalability trough parallelization

Semantic Data Integration for Early Clinical Development Use Case

- Improve the capability to integrate and interpret heterogeneous data
  - Evaluate hypotheses about patient characteristics and other factors that can explain segmentation criteria
  - Data interpretation is a non-trivial process that requires overcoming:
    - Semantic differences in the format, e.g. used identifiers
    - Verify, validate and compare experimental results with other established data sets
    - Efficient secondary usage of past experimental results and analysis conducted in later phases
  - Signal evaluation of adverse drug event reports
    - evaluate if there is a casual relationship between the drug and the adverse event (method RUCAM)

Our Objectives

- Integrate information using RDF data model
  - Integrated data sources to cover the path:

    gene – proteins – pathways – targets – disease – drugs – patient
- Reason over the integrated dataset
  - Remove redundancy / generate new links
  - Derive new implicit knowledge (e.g., “caspase activation via cytochrome c” is special form of “apoptosis regulation”)
- Apply information extraction algorithms and generate semantic annotations
  - Perform named-entity recognition and analyze some of the literals (e.g, the document texts)
  - Validate the new relations with the structured information
- Do it on a very large scale!

Our Approach

- Release early. Release often!
- LinedLifeData is an early prototype:
  - To host RDF knowledge base
  - Optimized for large scale and high-performance reasoning
  - Web interface and SPARQL endpoint to access the knowledge
  - Based on OWLIM semantic repository
- LifeSKIM is prototype to:
  - Recognize biomedical entities in text
  - Use semantic repository to store the information
  - Based on KIM platform
- We aim for constant interactions with the researchers!

Prototype: LinkedLifeData - PIKB

- Platform to automate the process:
  - Infrastructure to store and inferences
  - Transform the structured data sources to RDF
  - Provide web interface to access the data
- Currently operates over OWLIM semantic repository
- LinkedLifeData - PIKB statistics:
  - Number of statements: **1,159,857,602**
  - Number of explicit statements: **403,361,589**
  - Number of entities: **128,948,564**
- Publicly available at: http://www.LinkedLifeData.com

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

Linked Life Data

Semantic Data Integration Platform for Biomedical Domain

SPARQL SELECT Query

Append namespaces: --namespaces--

Query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX biopax2: <http://www.biopax.org/release/biopax-level2.owl#>
PREFIX gene: <urn:lsid:ncbi.nlm.nih.gov:ontology:>
PREFIX lifeskim: <urn:lsid:ontotext.com:lifeskim:>
SELECT distinct ?path ?gene ?name
WHERE {
        ?path biopax2:PATHWAY-COMPONENTS ?i .
        ?path biopax2:NAME ?path_name .
        ?i rdf:type biopax2:interaction ;
        biopax2:CONTROLLER [biopax2:PHYSICAL-ENTITY ?pe] .
        ?pe biopax2:COMPONENTS [biopax2:PHYSICAL-ENTITY ?pr] .
        ?pr biopax2:XREF ?x .
        ?x lifeskim:hasReferenceTo ?gene .
        ?gene gene:chromosome "7" ;
        gene:hasOfficialName ?name
} LIMIT 100
```

Contexts:

- BioGRID - General Repository for Interaction Datasets
- NCI - National Cancer research Institute
- BioCarta - Life science information provider for proteomics
- BioCyc - Description of metabolic pathways and their associated enzymes

Apply contexts

Predefined queries: --search topic-- | Select pathways controlled by expression of genes loca

Results limit: 100

- --Select query--
- SPARQL Select template
- Select interactions where participates specified protein
- Select interacting partners for specified protein
- Select equivalent interactions from different sources
- Select pathways controlled by expression of specified gene
- Select pathways controlled by expression of genes located on specified chromosome
- Select pathways associated with specified GO term
- Select biological processes controlled by expression of specified gene

OWLIM

ontotext

Linked Life Data

Semantic Data Integration Platform for Biomedical Domain

Explore Resource: urn:lsid:geneontology.org:go:GO:0005737

Perform another query

Statements with this resource as subject

Subject:	Predicate:	Object:
this	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://proton.semanticweb.org/2006/05/protons#Entity
this	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	urn:lsid:geneontology.org:ontology:GeneOntologyTerm
this	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	urn:lsid:geneontology.org:ontology:CellularComponent
this	http://proton.semanticweb.org/2006/05/protons#generatedBy	http://www.ontotext.com/kim/2006/05/wkb#Trusted_Instance
this	urn:lsid:geneontology.org:ontology:is_a	urn:lsid:geneontology.org:go:GO:0005575
this	urn:lsid:geneontology.org:ontology:is_a	urn:lsid:geneontology.org:go:GO:0044424
this	urn:lsid:geneontology.org:ontology:is_a	urn:lsid:geneontology.org:go:GO:0044464
this	urn:lsid:geneontology.org:ontology:hasDefinition	"All of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures."
this	http://proton.semanticweb.org/2006/05/protons#mainLabel	"cytoplasm"

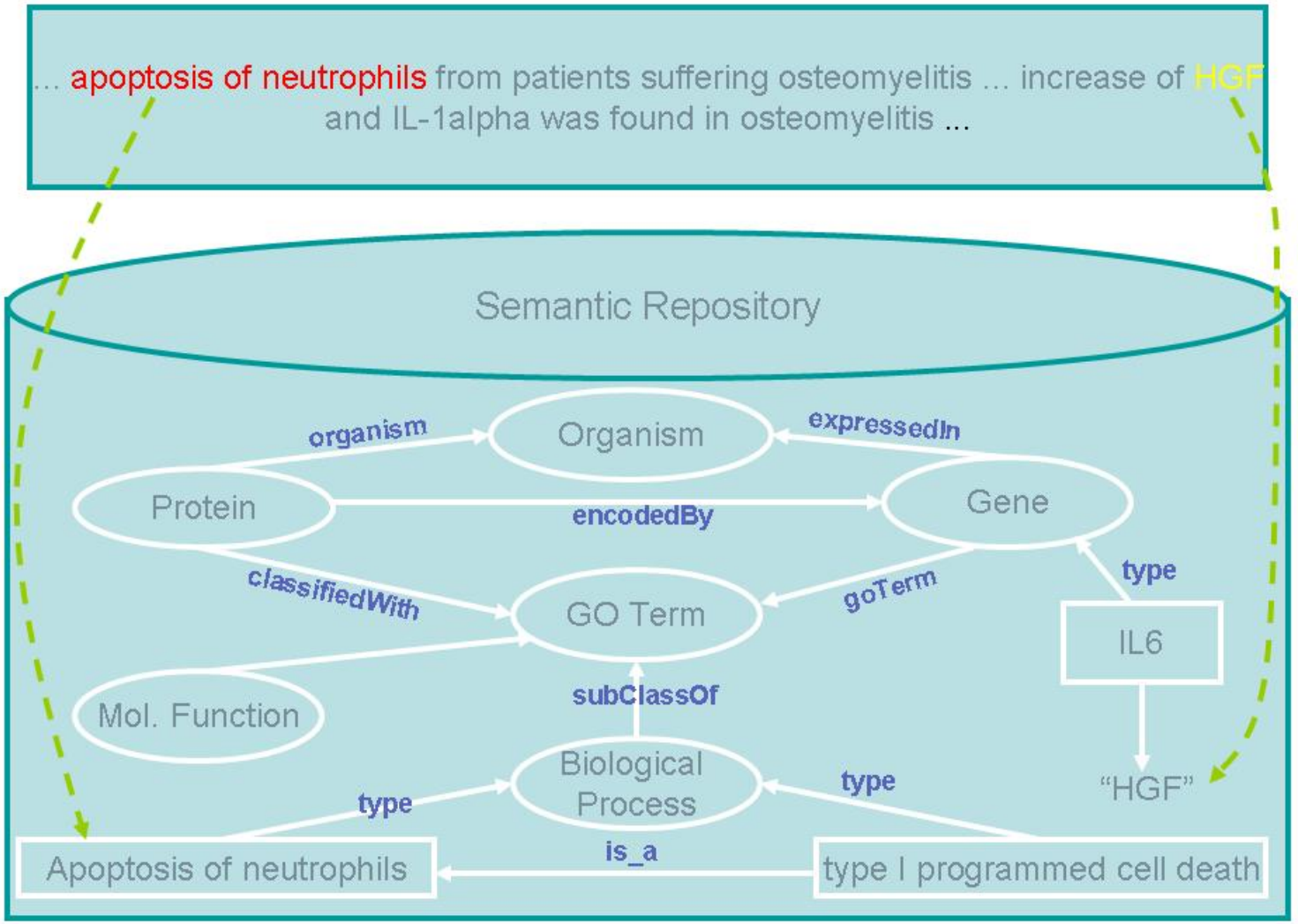
Statements with this resource as object

Subject:	Predicate:	Object:
urn:lsid:ncbi.nlm.nih.gov:gene:3258531	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:2542419	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:856021	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:84446	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:381979	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:54014	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:93871	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:244813	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:7809	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:8927	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:12217	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:54836	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:192120	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:112939	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:66830	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:53339	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:83962	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this
urn:lsid:ncbi.nlm.nih.gov:gene:55643	urn:lsid:ncbi.nlm.nih.gov:ontology:goTerm	this

Semantic Annotation Generation

- Semantic annotations stands for:
  - Textual references to formally described entity in ontology (e.g., LLD)
  - The process of semantic annotations generation
  - Closes the gap between the structured and unstructured knowledge
- Common tasks related to semantic annotations
  - Automatic semantic annotations of text
  - Ontology population
  - Semantic indexing and retrieval of content
  - Query and navigation involving structured knowledge
  - Combination with classical information retrieval tasks

Semantic Annotation



Current Entity Categories

- Gene names (Entrez-Gene)
- Gene and gene production annotations (Gene Ontology)
- Organisms (NCBI Taxonomy)
- Diseases (SNOMED from UMLS)
- Drug compounds (DrugBank)
- The classes Ambiguous gene, Cell Line, DNA and RNA are automatically learned from text

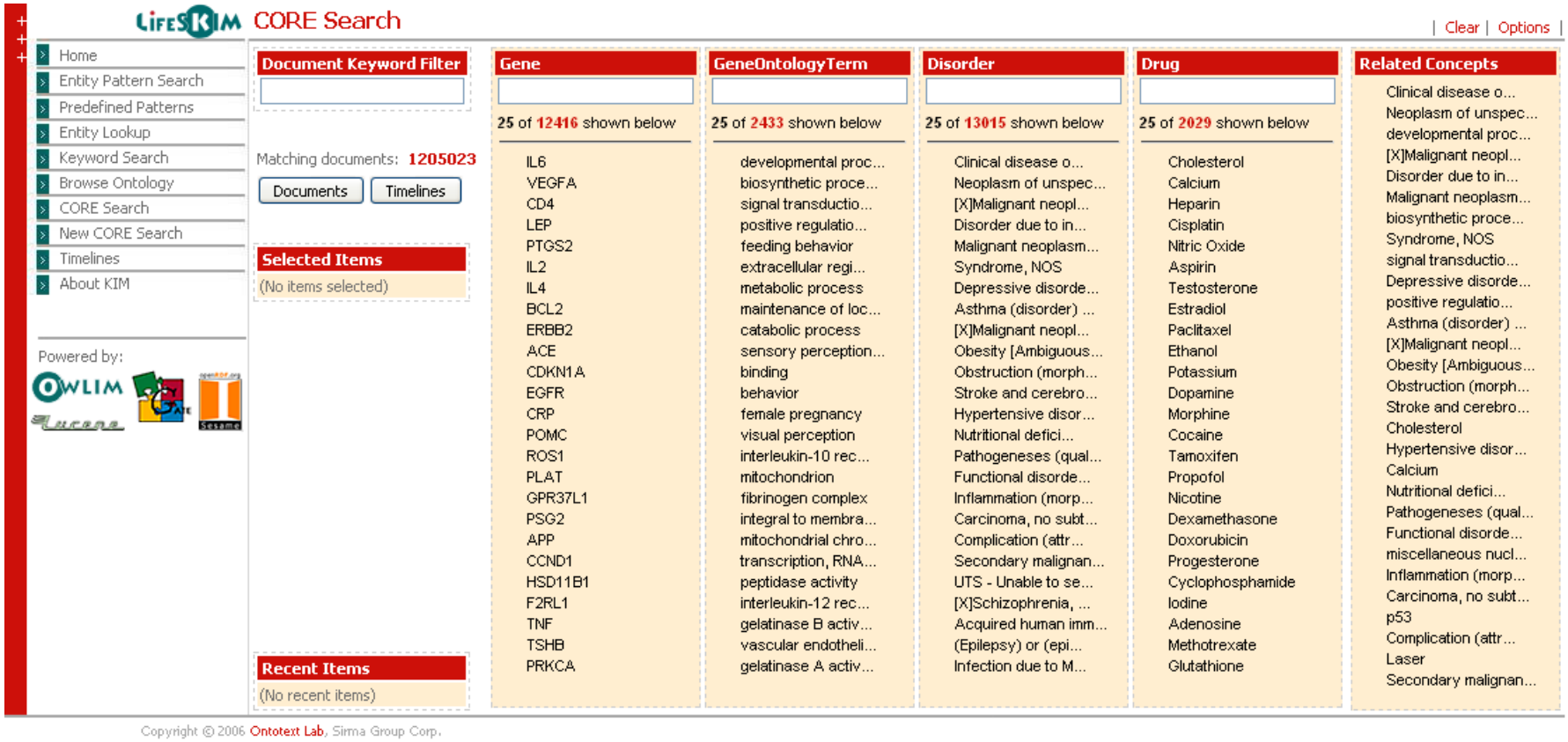
Results of the Semantic Annotation Process

- 1,204,063 Medline abstracts are annotated
- 10,884,032 semantic annotations are created
- Saved links to 40,510 existing entities

Type	
Genes	12,416
Organism	10,617
Diseases	9,256
Drugs	2,029
Neoplastic process	1,667
Biological process	1,604
Pathological functions	1,342
Mental/behaviour dysfunction	749
Molecular function	624
Cellular component	205
DNAs (newly recognized)	156,426
Cell lines (newly recognized)	89,217
Cell types (newly recognized)	85,199
RNAs (newly recognized)	6,001

Prototype: LifeSKIM

To get access to prototype send a request:
lifeskim@ontotext.com



Acknowledgement

- Ontotext
  - Deyan Peychev, Georgi Georgiev, OWLIM team
- AstraZeneca
  - Elisabet Söderhielm, all the scientists
- LarKC consortium / LarKC is partially funded by EU FP7-215535