ontotext
Semantic Technology Lab

Linked Life Data for annotation of Medline

Semantic data-integration and search in the life science domain

Vassil Momtchev (Ontotext)

ontotext
Semantic Technology Lab

Outline

- Life science and health care vertical – opportunity for semantic technology
- How RDF technology will help the end-user
- Linked Life Data – a platform for semantic data integration
- LifeSKIM – A smart textual analysis backed by an ontology

Innovation or Stagnation
What's the Diagnosis?

- Investment & progress in basic biomedical science has for surpassed investment and progress in the medical product development process
- The development process – the critical path to patients – becoming a serious bottleneck to delivery of new products
- We are using the evaluation tools and infrastructure of the last century to develop this century's advances

From FDA presentation on Critical Path for Science Board
by Janet Woodcock, 2004/04/26

Andy Law's First and Second Laws

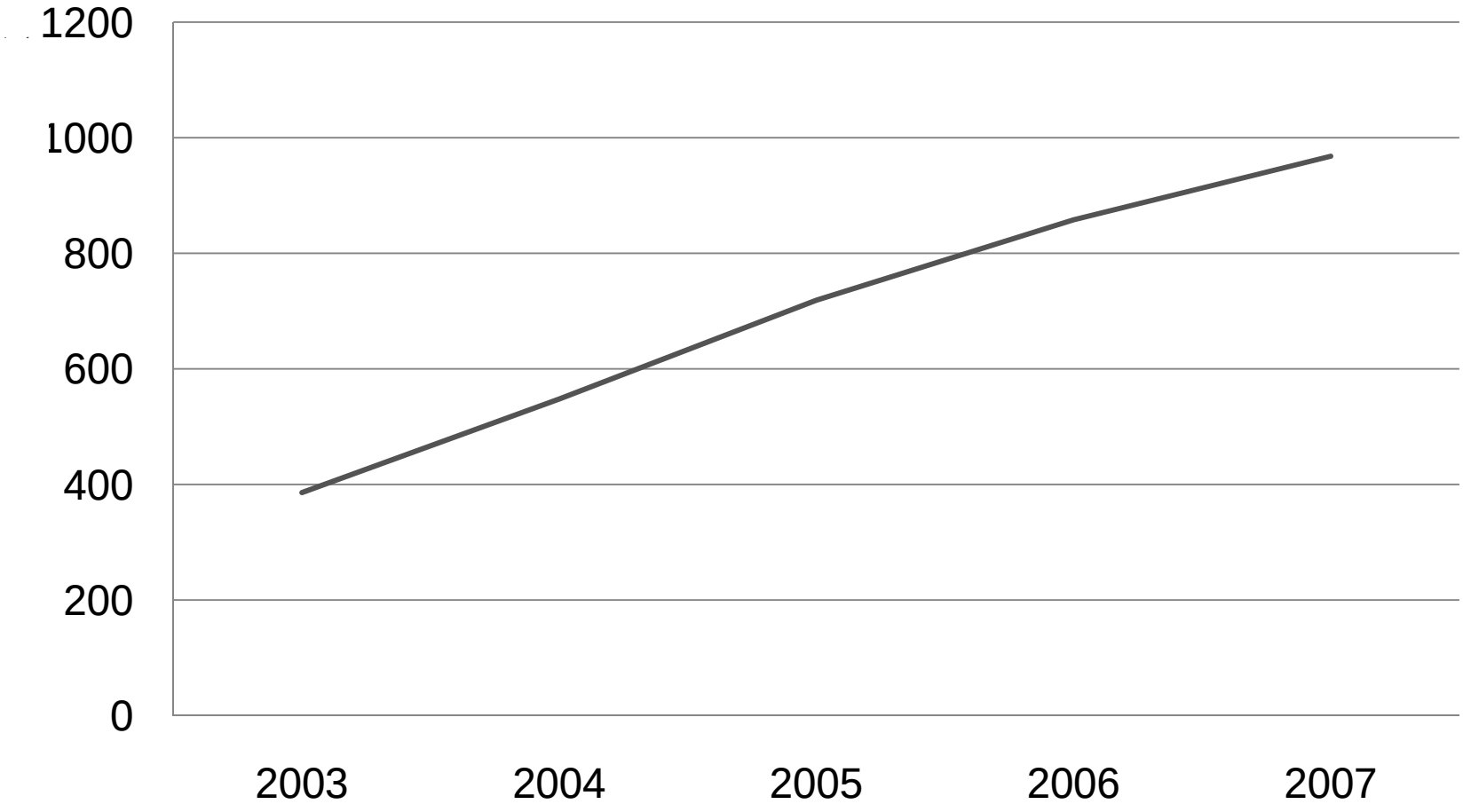
"The first step in developing a new genetic analysis algorithm is to decide how to make the input data file format different from __all__ pre-existing analysis data file input formats."

*"**The** second step in developing a new genetic analysis algorithm is to decide how to make the **output data file format** incompatible with __all__ pre-existing data file formats."*

ontotext
Semantic Technology Lab

TAO

Take Your Best Guess

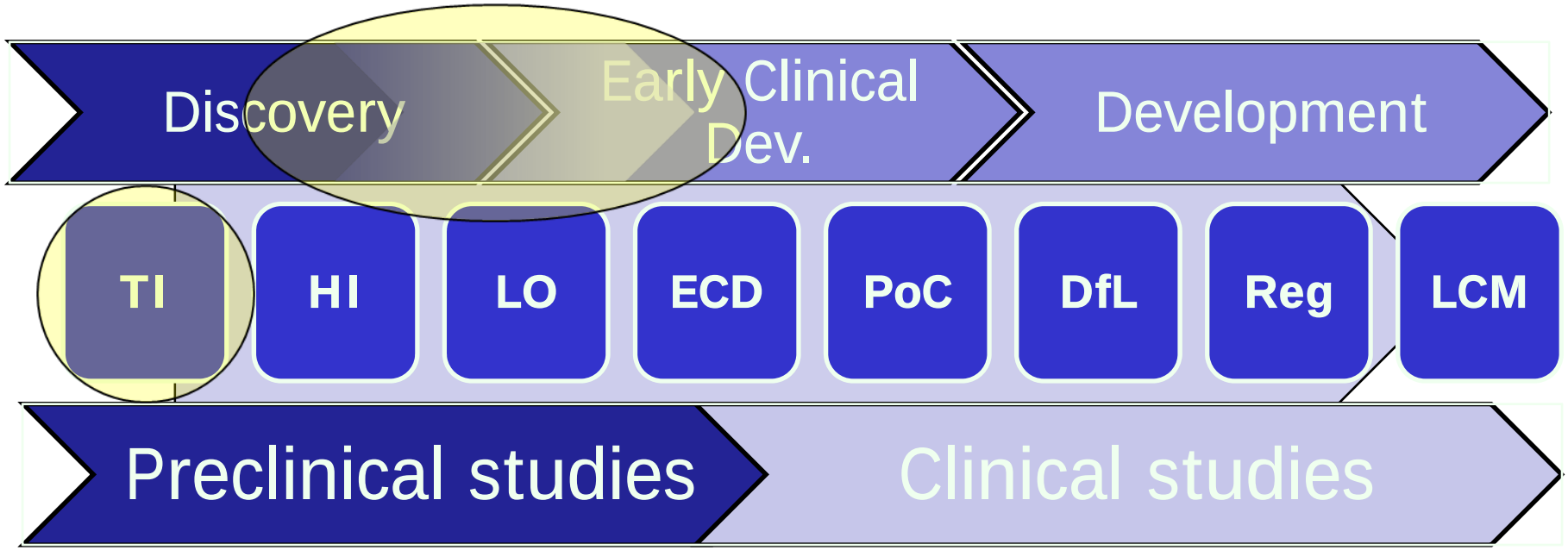


The Problems

- The data is supported by different organizations
- The information is highly distributed and redundant
- There are tons of flat file formats with special semantics
- The knowledge is locked in vast data silos
- There are many isolated communities which could not reach cross-domain understanding

Massive data integration and interpretation problem!

Drug Development Process



- Target Identification
- Hit Identification
- Lead Optimisation
- Proof of Concept
- Development for Launch
- Registration and Launch
- Life Cycle Management

The Questions in Early Clinical Development

The "*translation*" of basic research into real therapies for real patients – *Translational Medicine*

Understand the drug in context of:

- the disease
  - The chemistry/pharmacology process
  - How to measure?
  - What causes the disease?
  - How does the disease evolve?
- the patient
  - What different phenotypes exists?
  - Are there different Genetic profiles?

The Challenge

- Develop compound and knowledge to prove its target population
- Analyze the vast amounts of existing information
- A successful project lasts for 7 to 15 years

The Health Care and Life Science Industry Needs

- Support incremental extension of the knowledge base with highly heterogeneous data sets
- Allow straightforward updates of the information
- Provide scientists with computational support to conceptualize the breath and depth of relationships between data
- Analyze unstructured information

The need of powerful heterogeneous knowledge stores

Which Technology to Choose?

Possible Solutions

Classical data-integration with:

- Data warehouses
- Federa
  framev
- Databa
  techno

Not really...

- Mapping works efficiently on a small scale

paradigm

llenge

sually does

to integrate textual information

We are using the evaluation tools and infrastructure of the last century to develop this century's advances

Semantic Data Integration Benefits

- To overcome the different semantic and syntax representation
- To handle inconsistencies problems related to incomplete data or different versions
- To unlock the data stored in silos and solve container-reference dichotomy – data once stored and connected is hard to rearrange and connect in new ways

- How semantic web technology could help to end users?

What is Semantic Web?

Enrich the existing web

- Recipe:
  - Annotate, classify, index
- Meta-data from:
  - Automatically producing mark-up: named-entity recognition concept extraction, tagging, etc.
- Enable personalisation, search, browse...

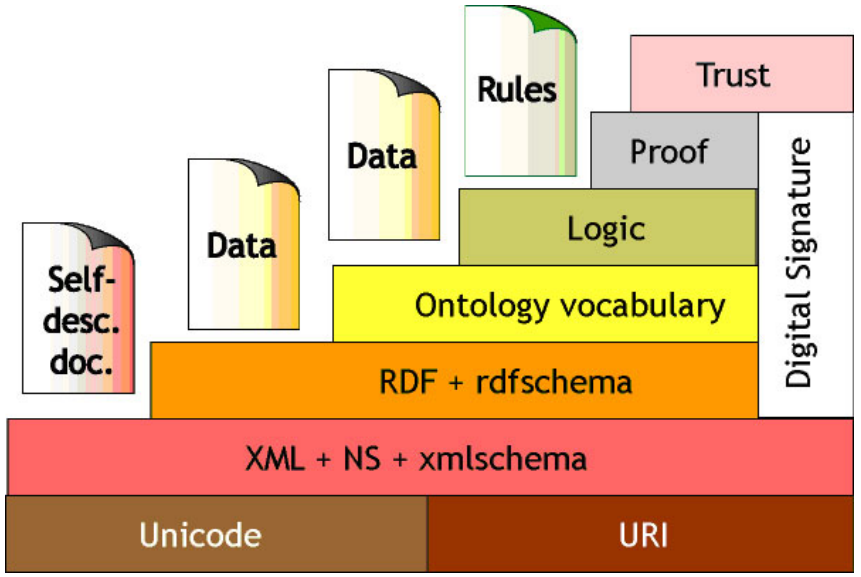
Semantic Web as Web of Data

- Recipe:
  - Expose data on the web, use RDF, integrate
- Meta-data from:
  - Expressing DB schema semantics in machine interpretable ways
- Enable integration and unexpected reuse

Source: Frank van Harmelen RDF presentation

W3C Stack

- XML
  - Surface syntax, no semantics
- XML Schema
  - Describes structure of XML documents
- RDF
  - Data model for “relations” between “things”
- RDF Schema
  - RDF Vocabulary Definition Language



The picture is a bit out-dated today

So Why No Just Use XML?

```
<country name=”Sweden”>
    <capital name=”Stockholm”>
    <areacode>01</areacode>
    </capital>
</country>
```

```
<nation>
    <name>Sweden</name>
    <capital>Stockholm</capital>
    <capital_areacode>01
    </capital_areacode>
</nation>
```

No agreement on**:**
Structure
- is country a:
  - object?
  - class?
  - attribute?
  - relation?
  - something else?
- what nesting mean?

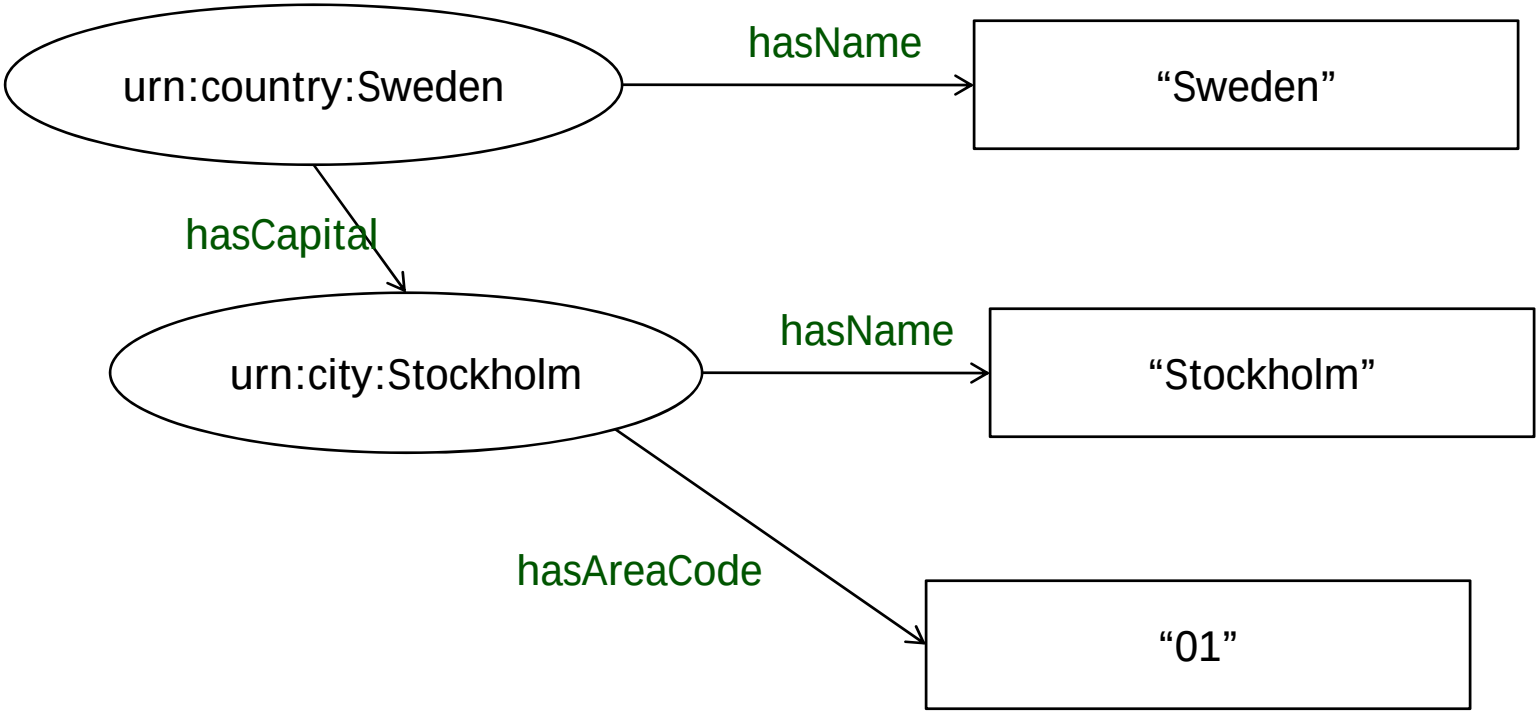
Vocabulary
- is country same as nation?

Are the above XML documents the same?
Do they convey the same information?
Is that information machine-accessible?

What is RDF?

- RDF
  - stands for Resource Description Framework
  - is a W3C Recommendation (http://www.w3.org/RDF)
- RDF is a data model
  - for representing meta-data (data about data)
  - for describing the semantics of information in a machine-accessible way
- What can you use it for?
  - intelligent information brokering
  - meaning-based computing
  - agent communication

How RDF looks like?

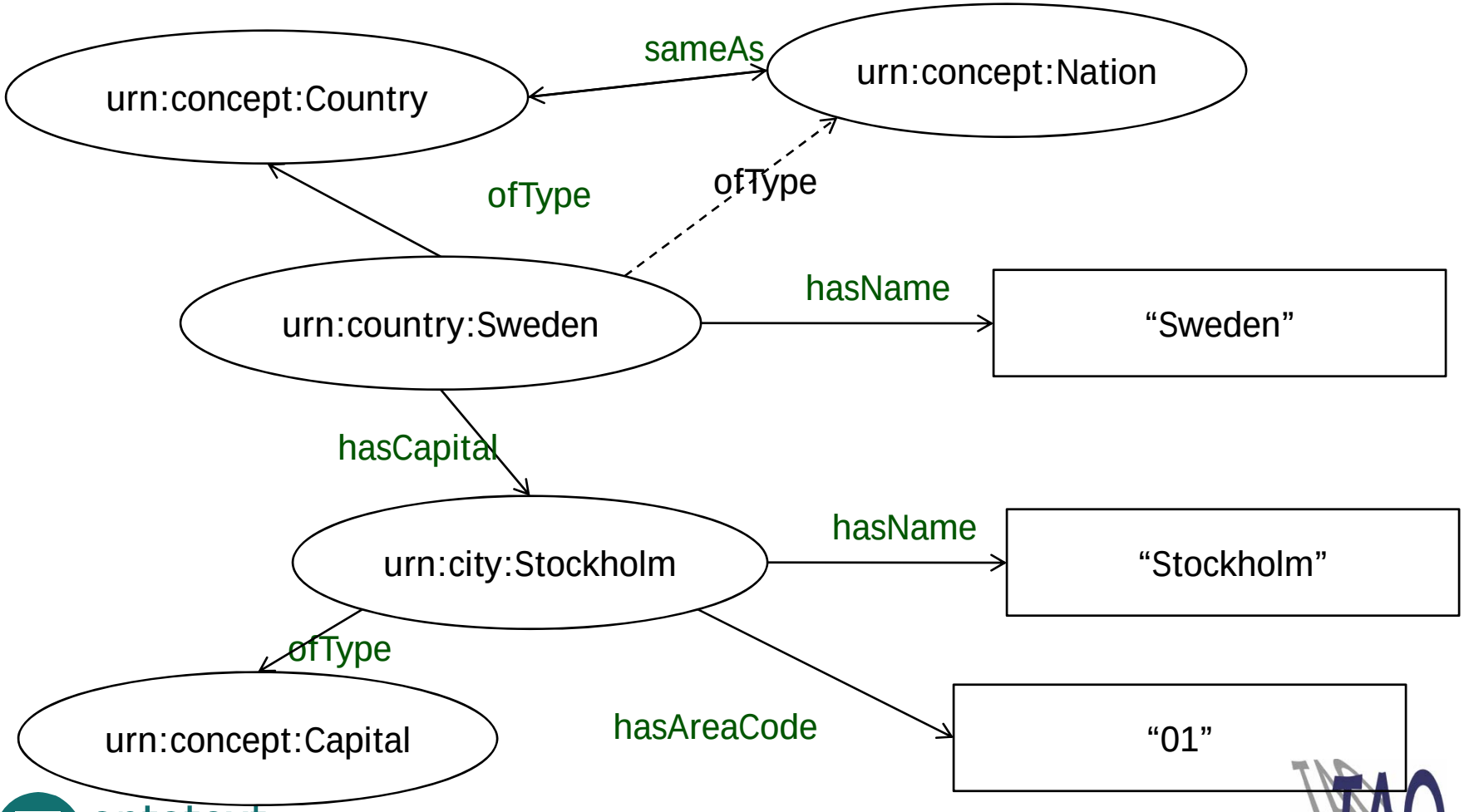


Subject	Predicate	Object
urn:country:Sweden	hasName	“Sweden”.
urn:country:Sweden	hasCapital	urn:city:Stockholm.
urn:city:Stockholm	hasName	“Stockholm”.
urn:city:Stockholm	hasAreaCode	“01”.

ontotext
Semantic Technology Lab

TAO

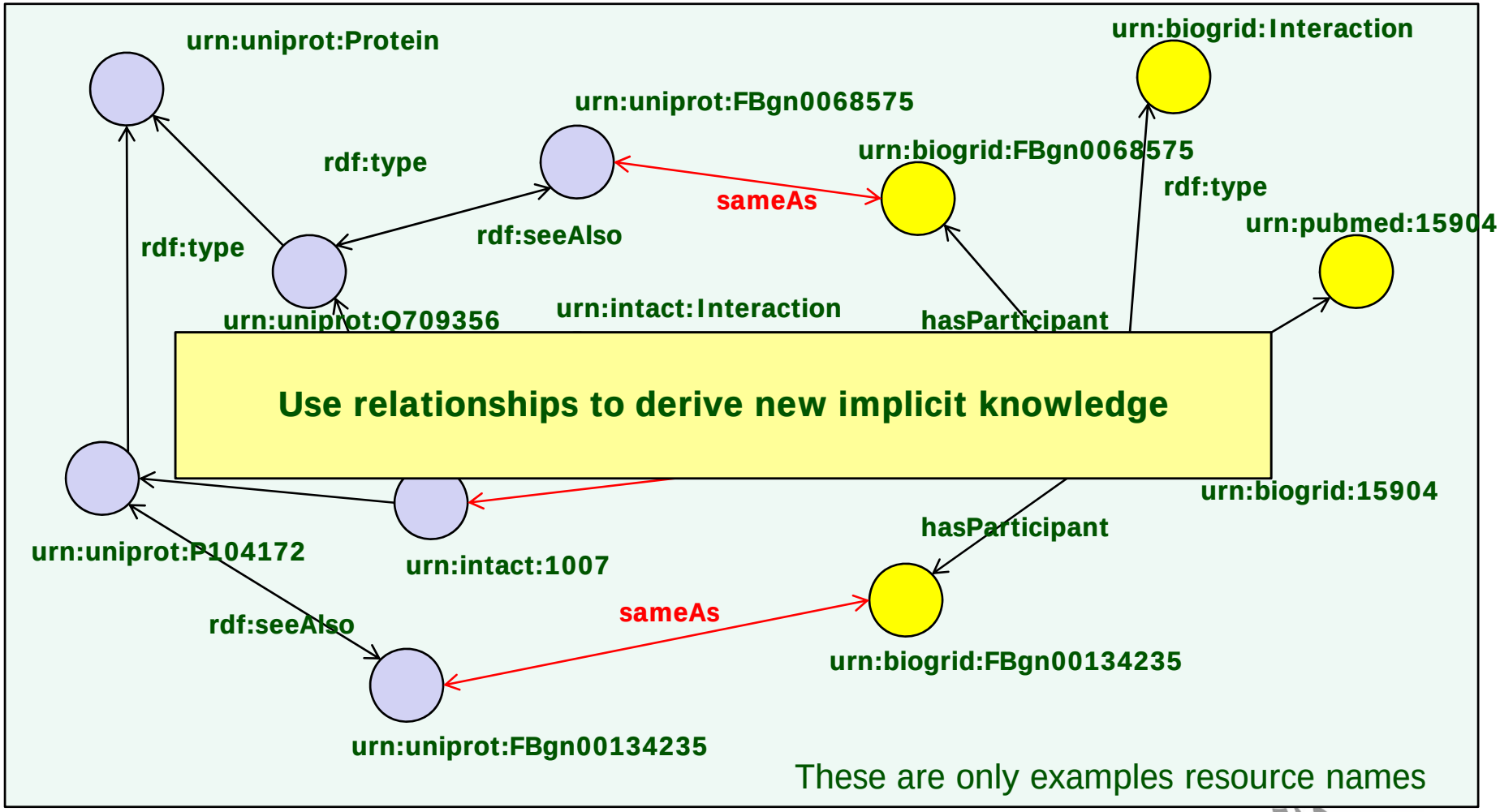
RDF Schema and further interpretation



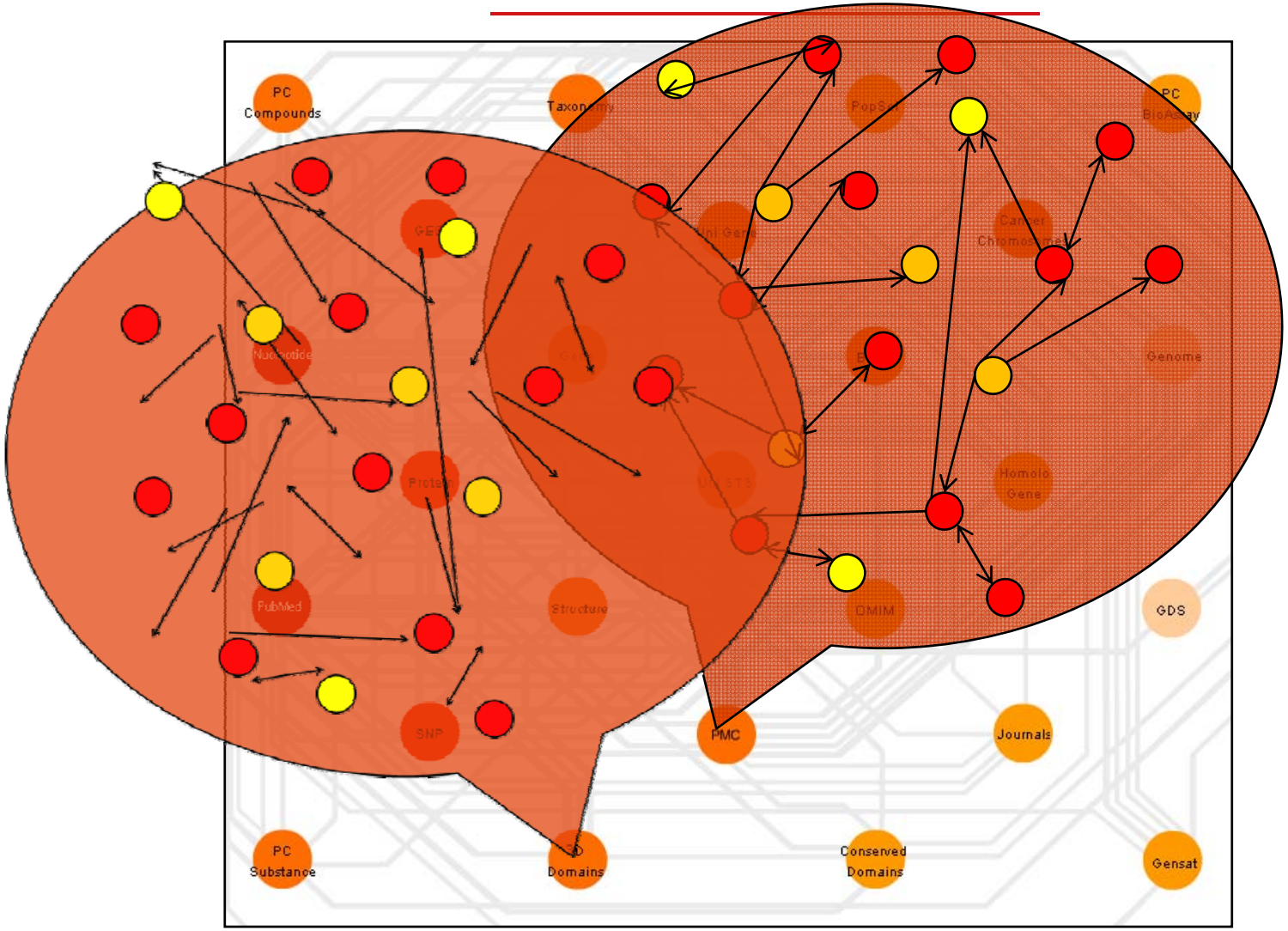
ontotext
Semantic Technology Lab

TAO

RDF for Life Sciences



Entrez Databases



Linked Life Data

- Linked Life Data stands for a platform to:
- Operate with heterogeneous data sets
- Allow semantic data integration
- Provide tools for knowledge access and management
- Compliant with W3C standards and recommendations
- Developed in collaboration with AstraZeneca in LarKC project

Our Objectives

- Integrate the linked 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”)
- Do it on a very large scale!

Data Sources

Ty

-
-
-
-
-
-
-

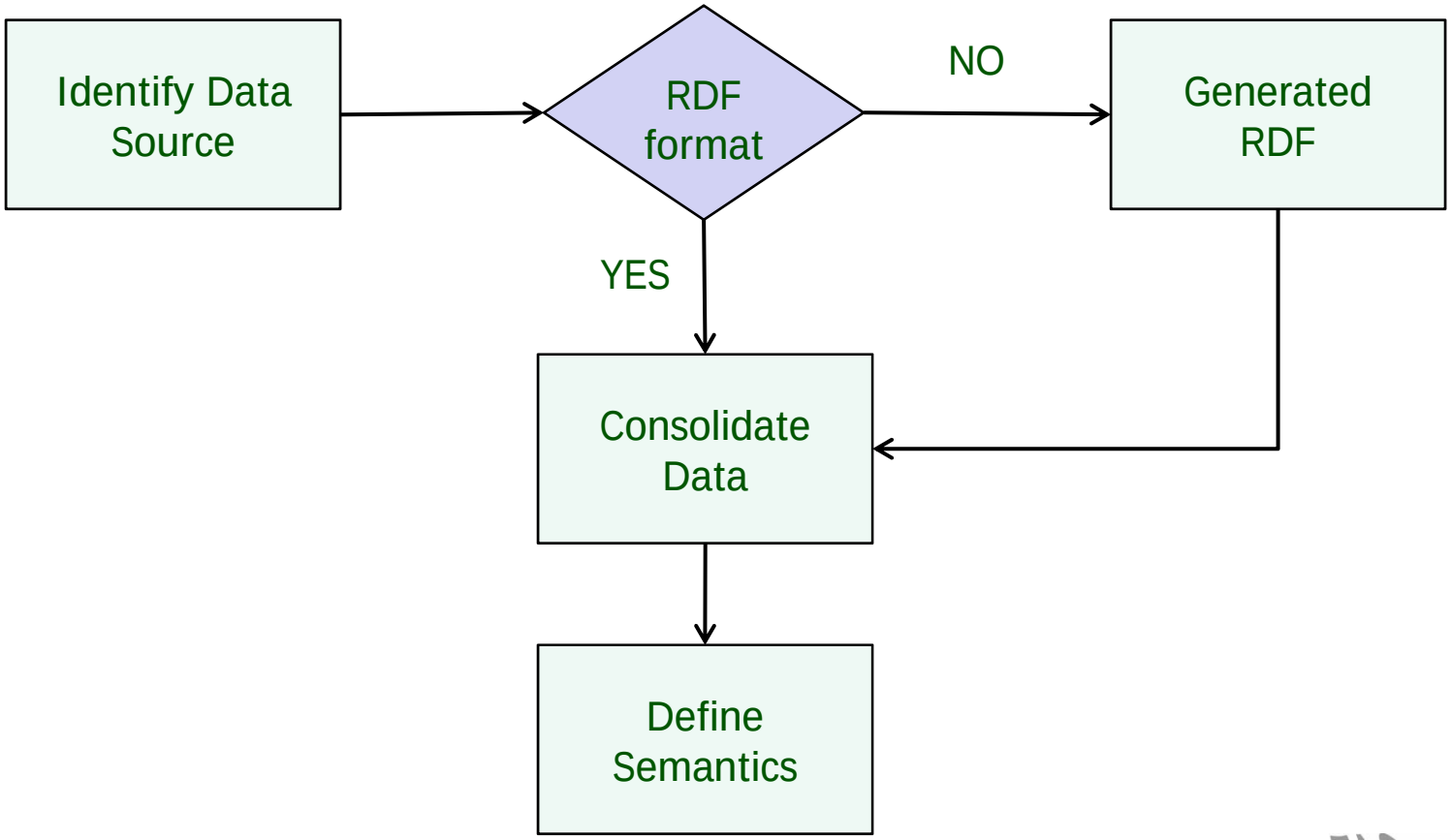
Sometimes we need to ask far more questions efficiently:

Give me all proteins which interacts in nucleus and are annotated with repressor and have at least one participants that is encoded by gene annotated with specific term and is located in chromosome X? Filter the results for Mammalia organisms!

pathways

BioCarta, KEGG, BioCyc

The Approach



Challenges to Overcome

- Syntactic
  - The way the different are serialized
- Structure
  - The way the different entities are represented
- Semantic
  - The way the different entities are interpreted

- W3C standard serialization formats for data exchange
- The graph model used by RDF gives maximum flexibility
- Support custom R-entailment rules to derive meaning

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 Overview

- Platform to automate the process:
  - Infrastructure to store and inferences
  - Transform the structured data sources to RDF
  - Provide web interface and SPARQL endpoint to access the data
- Currently operates over OWLIM semantic repository
- Linked Life Data statistics:
  - **gene – proteins – pathways** – targets – disease – drugs – patient
  - 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

Linked Life Data

Semantic integration of biological databases

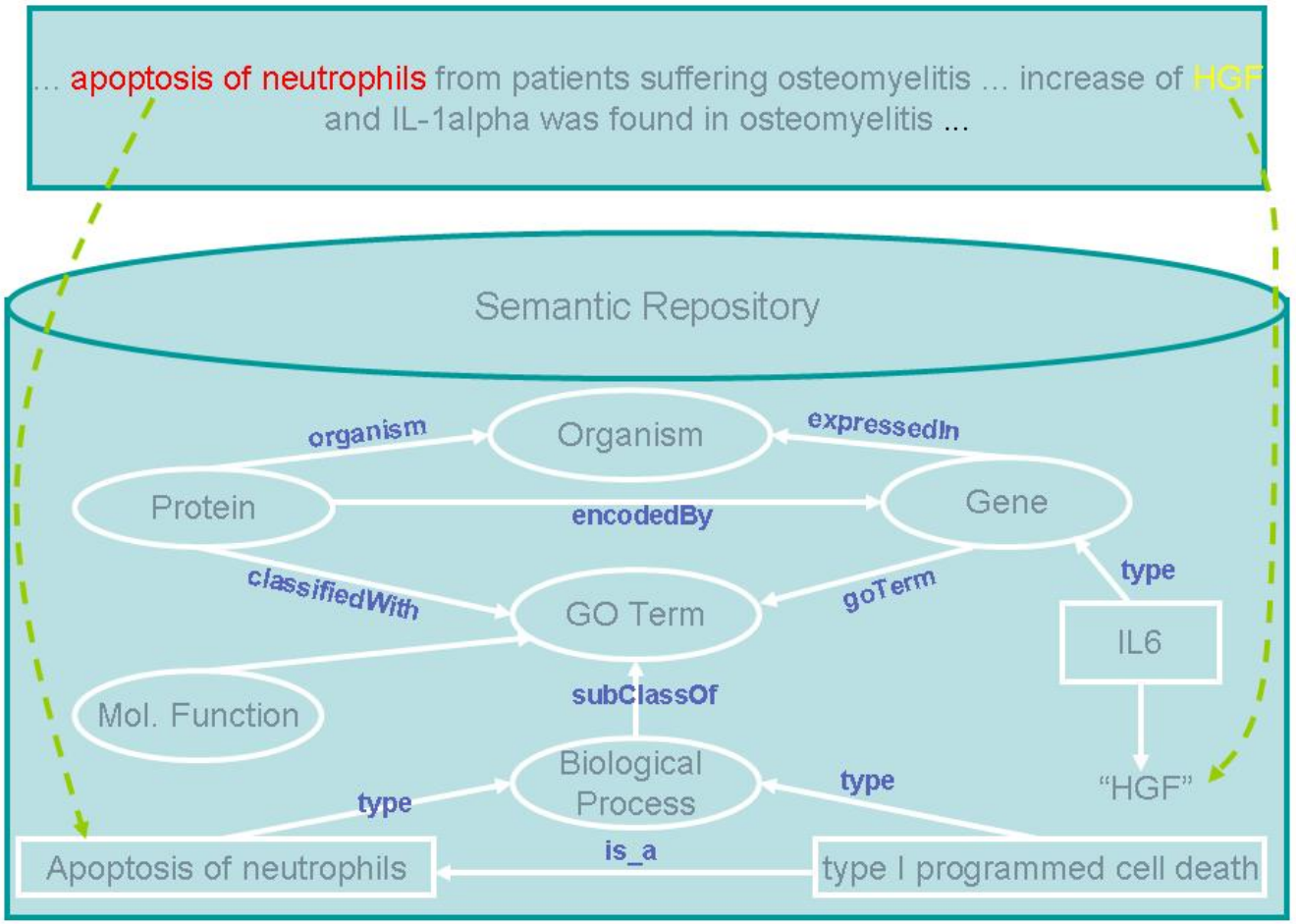
ontotext
Semantic Technology Lab

The way to semantic Service Oriented Architecture

LifeSKIM – Quick Facts

- LifeSKIM application provides a scalable support of:
- Querying and navigation of knowledge generated from structured (biological databases) and unstructured (biomedical document);
- Semantic indexing and retrieval of document using ontology
- Ontology population and learning of new types of entities from text
- Efficient reasoning against the extracted and structured information, e.g., “type I programmed cell death” is “Apoptosis of neutrophils” and “biological process” ;
- Co-occurrence and ranking of entities

Semantic Annotation Example



How LifeSKIM Searchers Better?

The classical IR could not match:

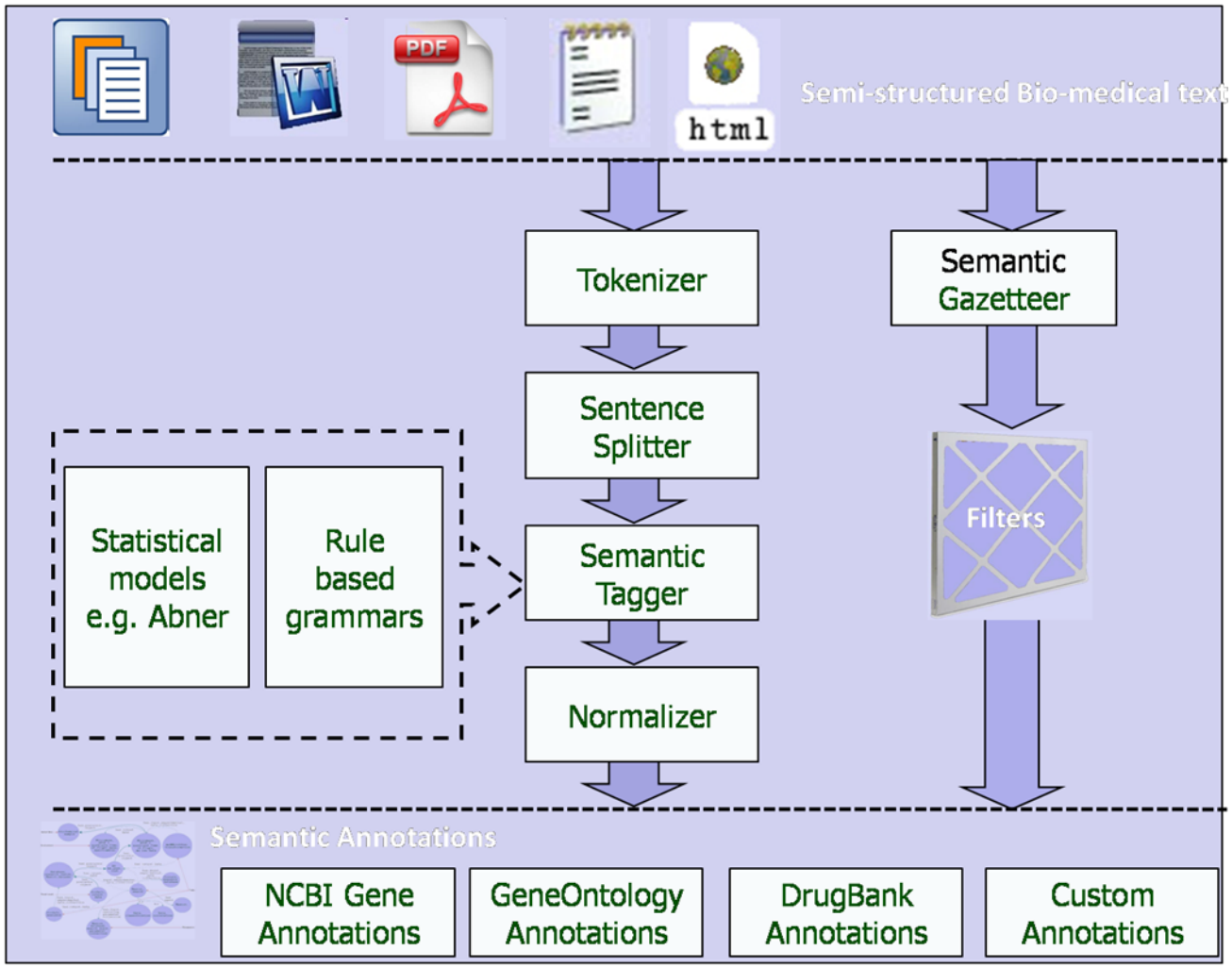
- interleukin 6 with a HGF or HSF or BSF2 or IL-6 or IFNB2

Interleukin 6 is a an entity in Entrez-Gene with GeneID: 3569, and HGF; HSF; BSF2; IL-6; IFNB2 are aliases for the same gene entity.

- apoptosis of neutrophils with “programmed cell death”;

GeneOntology thesaurus adds the above list of terms as part of apoptosis of neutrophils term.

A Complex IE Pipeline is Requred



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

LifeSKIM

Semantic annotation of biomedical documents