

Building a Pharmaceutical Competitive Landscape using Semantic Technologies

**Ranga Chandra Gudivada, PhD
Biomedical Informatics Scientist
Discovery Informatics
Eli Lilly & Company
Indianapolis
Indiana 46285
USA**

The Eli Lilly logo, featuring the word "Lilly" in a red, cursive script font.

Index

- **What is Competitive Landscape?**
- **What kind of questions do we ask for Competitive Landscape?**
- **Why “Semantics” to Competitive Landscape?**
- **Objectives & Success Factors**
 - **Why we need dictionaries?**
 - **Why NLP and Semantic Integration?**
 - **Why Knowledge Representation in RDF & OWL?**
- **Workflow**
- **Sample Queries (Simple – Moderately Complex – Complex)**
- **What are the key lessons and risks learned with this project?**
- **Overall Conclusion**

What is Competitive Landscape?

Competitive Landscape is to build a Competitive Intelligence (CI) infrastructure for purposeful, ethical and co-coordinated monitoring of the competitors in any industry within a specific market place to:

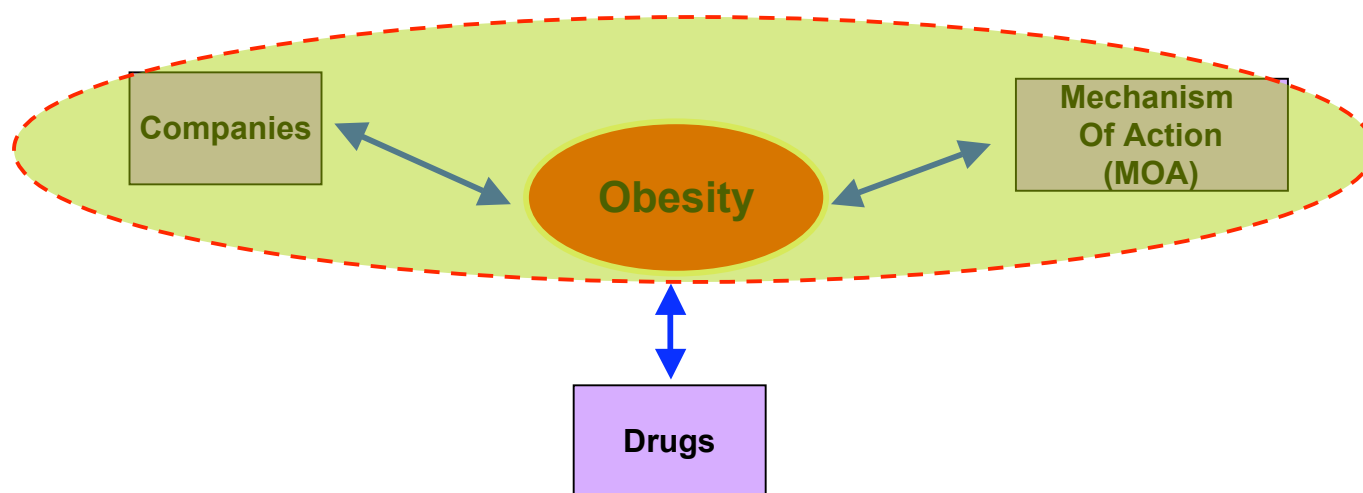
- Strategically gain foreknowledge of recent developments of your competitor's plans
- Make calculated informed business decisions and formulate operational strategy

Competitive Landscape @ Lilly

Endocrine PI project is initiated to provide a mechanism to actively survey the public information for competitive intelligence on Endocrine area and also provide easy access to the information to enable scientist to better position Lilly IP over other competitors

What kind of Questions do we ask for Competitive Intelligence?

- Which companies have drugs in phase-2 to treat obesity?
- What are the compounds for obesity and what are the parent companies for these compounds?
- Is there any company using GPCR modulators for obesity?
- Which companies are working on this particular MOA and on obesity?



Questions to the team before initiating project execution?

- Does in general Competitive Intelligence requires “Semantic” Component?
- Does Endocrine PI requires “Semantic Integration”?
- Are there any existing ontologies
 - Company Ontology
 - Mechanism of Actions (MOA) Ontology
- Does NLP or Text Mining methods work for this kind of data?
- Can inference and reasoning play a part to discover “buried” knowledge?
- Can Semantic Web Standards be applied for this real practical application?

Why “Semantics” to Competitive Intelligence

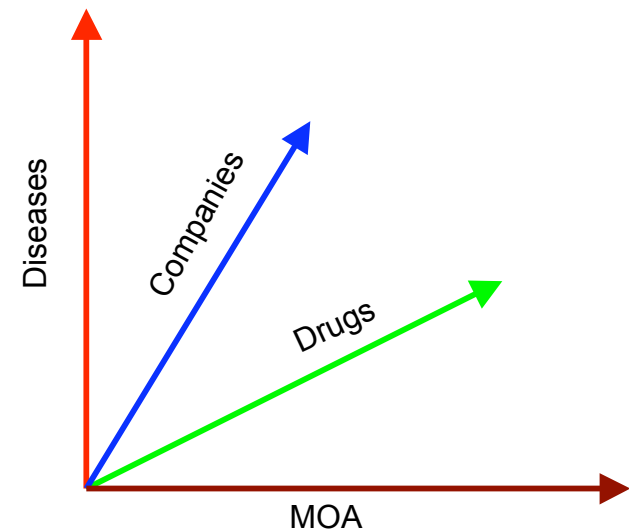
At Macro - Level

- Information Integration
- Analysis
- Inference & Reasoning
- Effective Search & Retrieval
- Creating a Knowledge Infrastructure

At Micro - Level

- Multi – dimensional data
- Semantic variations
- Syntactic variations
- Unstructured & noisy text
- Subsumption (Parent – child) Relations

Multi – Dimensional knowledge
rich Information



Why 'Semantics' in Competitive Intelligence (Examples)

Syntactic Variations

Company	<ul style="list-style-type: none"> ➤ Merck & Co ➤ Merck & Co Inc ➤ Merck ➤ Merck & Co Ltd
MOA	<ul style="list-style-type: none"> ➤ Alpha-glucosidase inhibitor ➤ Glucosidase inhibitor alpha ➤ IGF binding protein-3 stimulator ➤ IGF binding protein stimulator-3

Semantic Variations

Company	<ul style="list-style-type: none"> ➤ Amgen Boulder Inc ➤ Applied Molecular Genetics Inc ➤ Synergen Inc ➤ Amgen
MOA	<ul style="list-style-type: none"> ➤ Serotonin 2A receptor antagonists ➤ 5-HT 2 receptor antagonist ➤ 5-HT2a antagonist ➤ Peroxisome proliferator-activated receptor delta antagonist ➤ PPAR delta antagonist ➤ Melanin concentrating hormone receptor 1 antagonists ➤ MCH receptor-1 antagonist ➤ Melanin concentrating hormone receptor 1 antagonists ➤ G-protein coupled receptor-24 antagonist

Parent – Child Relations

- ▼ Eli_Lilly_Co
 - Applied_Molecular_Evolution_Inc
 - Beiersdorf_Lilly_GmbH
 - Chugai_Lilly_Clinical_Research_Co_Ltd
 - CPI_Cardiologische_Geraete_GmbH
 - CPI_del_Caribe_Ltd
 - CPI_Europa_BV
 - Delta_Holdings_Inc
 - Dista_Mexicana_SA_de_CV
 - Dista_Products_Ltd
 - Dow_Elanco

Company
Trees

- ▼ Transporter_protein_modulator
 - ▼ Adenine_nucleotide_translocator_modulator
 - ▶ Adenine_nucleotide_translocator_1_modulator
 - ▶ Adenine_nucleotide_translocator_2_modulator
 - ▶ Adenine_nucleotide_translocator_3_modulator
 - ▶ Adenine_nucleotide_translocator_inhibitor
 - ▼ Adenine_nucleotide_translocator_stimulator
 - Adenine_nucleotide_translocator_1_stimulator
 - Adenine_nucleotide_translocator_2_stimulator
 - Adenine_nucleotide_translocator_3_stimulator
 - ▼ ATP_binding_cassette_modulator

MOA
Trees

Endocrine PI – Objectives & Success Factors

Objective 1 : Ontology Generation

Generate an Ontology (in appropriate format from vendor XML files) for Companies and Mechanism of Action's (MOA) that act like a,

- Dictionary or Vocabulary
- Taxonomy (Parent – Child Tree)

Objective 2 : NLP & Semantic Integration

Mapping of raw Endocrine – PI data to Company and MOA Ontology's created in Objective-1 by applying Natural Language Principles (NLP) techniques

Objective 3 : Knowledge Representation in RDF

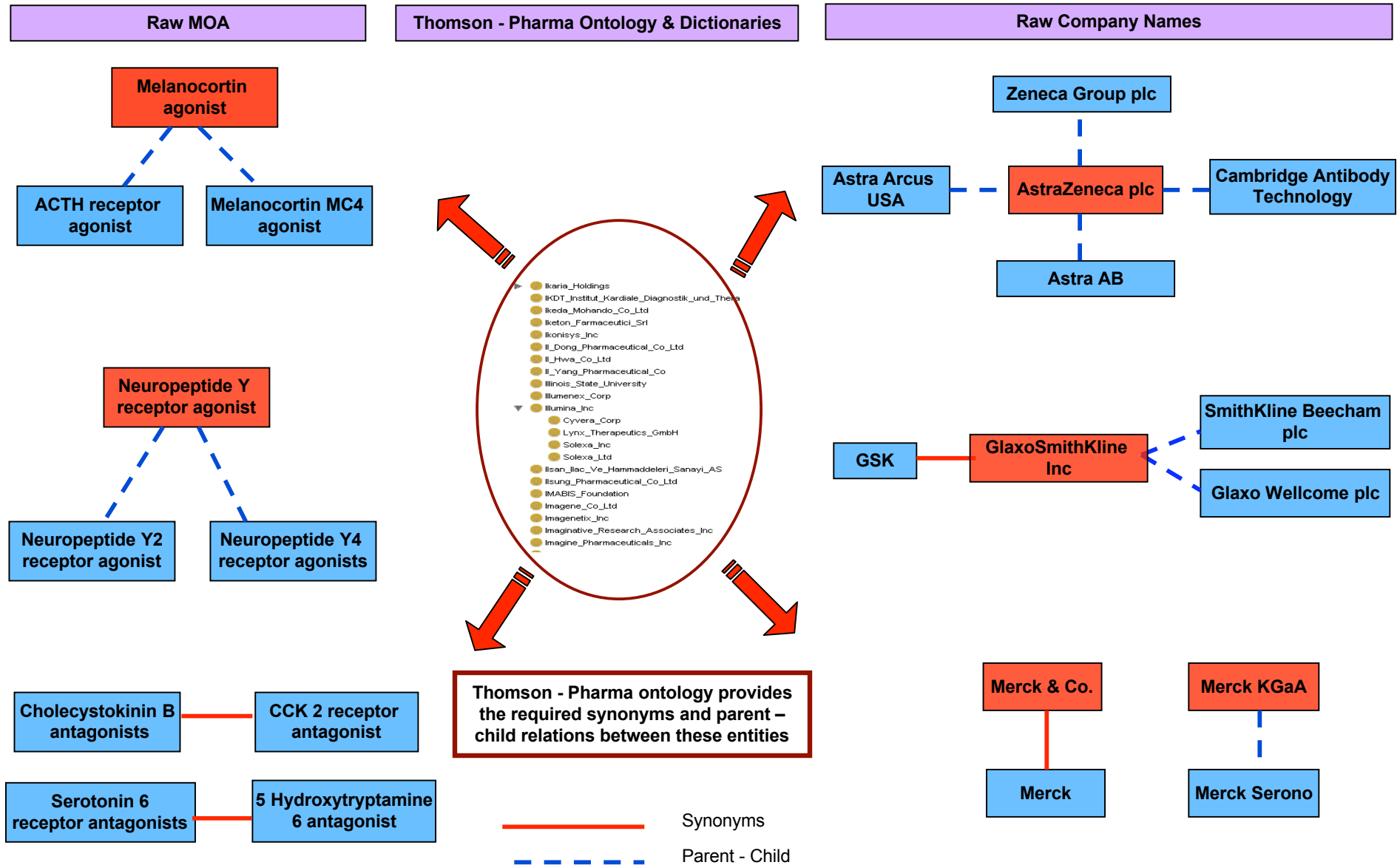
Representing semantically mapped Endocrine –PI raw data using Semantic Web standards (RDF & OWL)

Objective 4: Querying - Inference - Reasoning

Inferential querying of Endocrine -PI Competitive Intelligence information space (in RDF) to Investigate the advantages of Semantic Integration

Objective -1 : Ontology Generation

Why we need dictionaries?



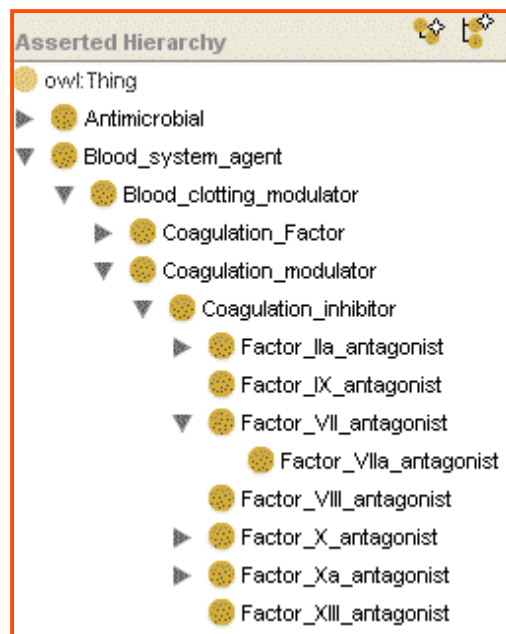
How did we get the Ontologies and Dictionaries?

Thomson – Pharma MOA in XML

```

- <ActionTree>
- <Action>
  <ActionName>Antimicrobial</ActionName>
  <TreeCode>ANT</TreeCode>
- <DrugList>
- <Drug>
  <DrugName>21g792 (anticancer), wyeth</DrugName>
  <Synonym>21g792</Synonym>
</Drug>
- <Drug>
  <DrugName>ad-1903</DrugName>
</Drug>
- <Drug>
  <DrugName>am-1939</DrugName>
  <Synonym>fluoroquinolones, kyorin pharmaceuticals</Synonym>
</Drug>
- <Drug>
  <DrugName>amoxicillin + clavulanate, gsk</DrugName>
  <Synonym>augmentin</Synonym>
</Drug>
- <Drug>
  <DrugName>ampicillin</DrugName>
  
```

From XML
To OWL



16300 Companies

Thomson – Pharma Company in XML

```

<?xml version="1.0" encoding="UTF-8" ?>
- <CompanyList>
- <Company ID="CM13594" CDLID="13594" timestamp="200803
  <CompanyDetail>
  - <CompanyCitation>
  - <CompanyName>
    <![CDATA[ Eppendorf-5 Prime Inc ]]>
  </CompanyName>
  <AddedDate>20020131164909</AddedDate>
  <UpdateDate>20030423111813</UpdateDate>
  - <RecordCount>
  <TotalCount>1</TotalCount>
  <DrugCount>0</DrugCount>
  <PatentCount>1</PatentCount>
  <HasAssociatedData>Y</HasAssociatedData>
  </RecordCount>
  - <AbbreviatedName>
  <![CDATA[ Eppendorf-5 Prime ]]>
  </AbbreviatedName>
  - <CompanySynonymList>
  - <Synonym CDLID="29041">
  - <SynonymName>
    <![CDATA[ 5 Prime ->3 Prime Inc ]]>
  </SynonymName>
  <IsDisplayName>N</IsDisplayName>
  </Synonym>
  - <Synonym CDLID="13594">
  - <SynonymName>
    <![CDATA[ Eppendorf-5 Prime Inc ]]>
  </SynonymName>
  <IsDisplayName>Y</IsDisplayName>
  
```

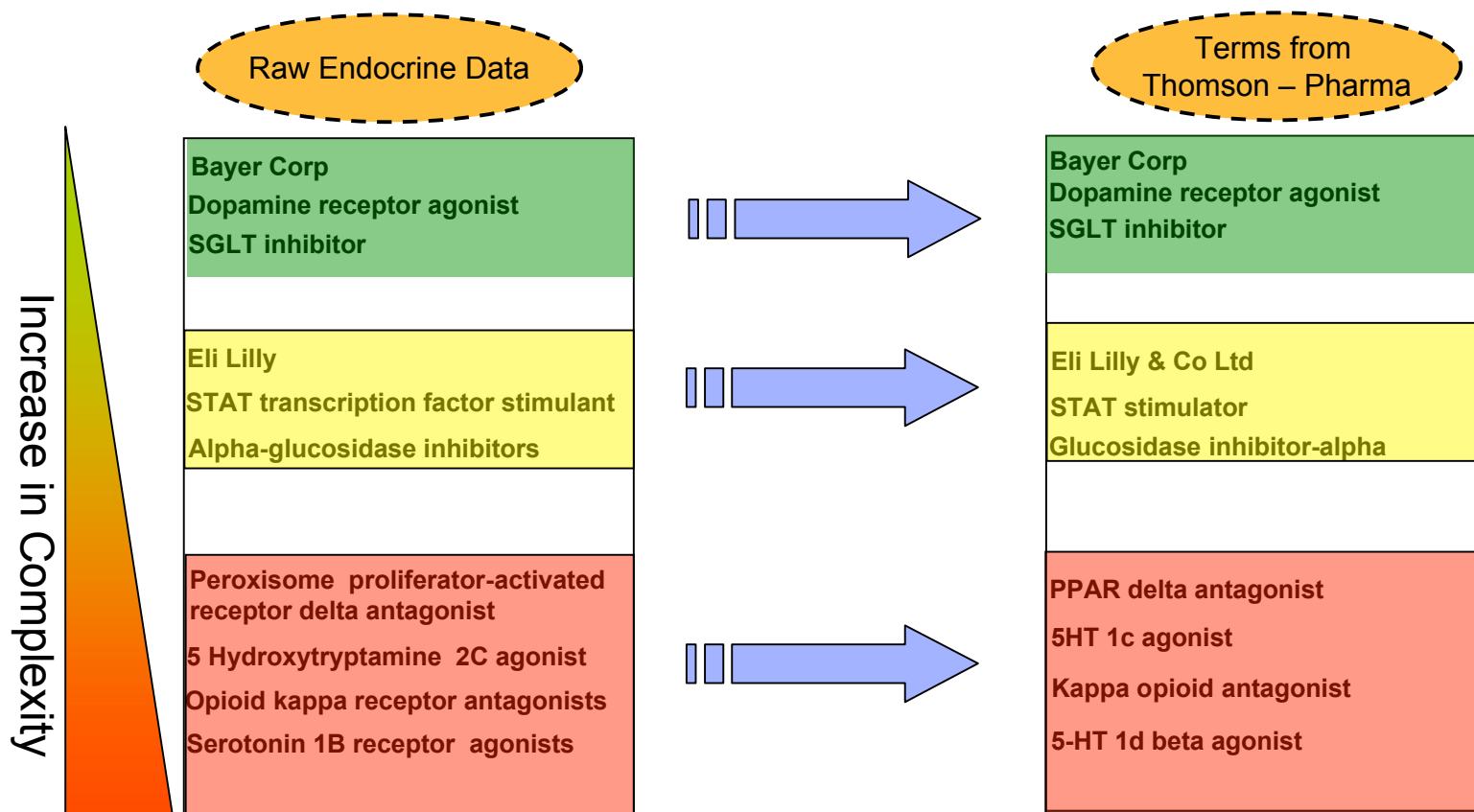
From XML
To OWL



16200 Companies

Objective 2 : NLP & Semantic Integration (Crucial Component)

Why NLP and Semantic Integration?



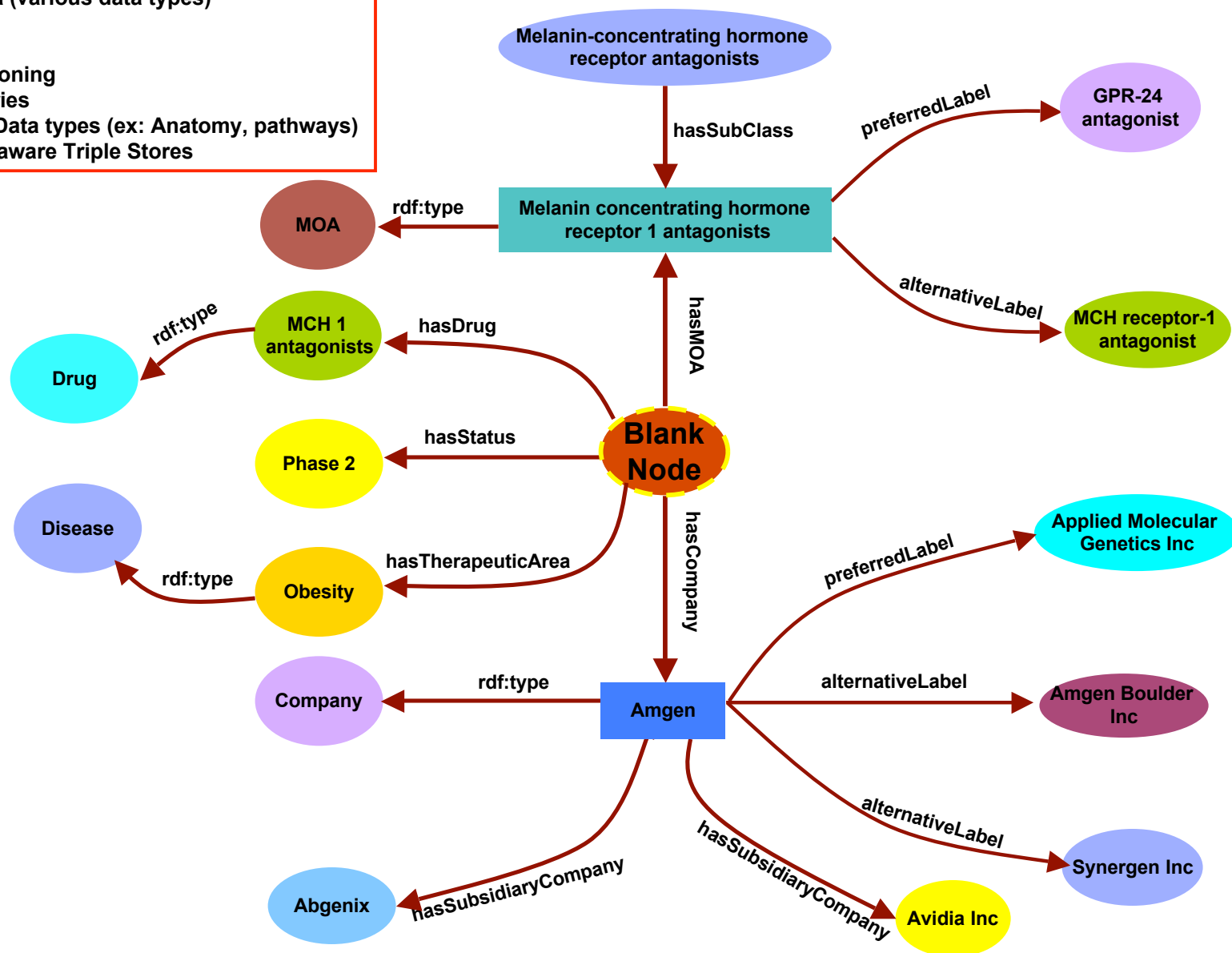
Raw endocrine data is gathered from

- Thomson Pharma
- Adis
- IMS
- PharmaProjects

Objective 3 : Knowledge Representation in RDF & OWL

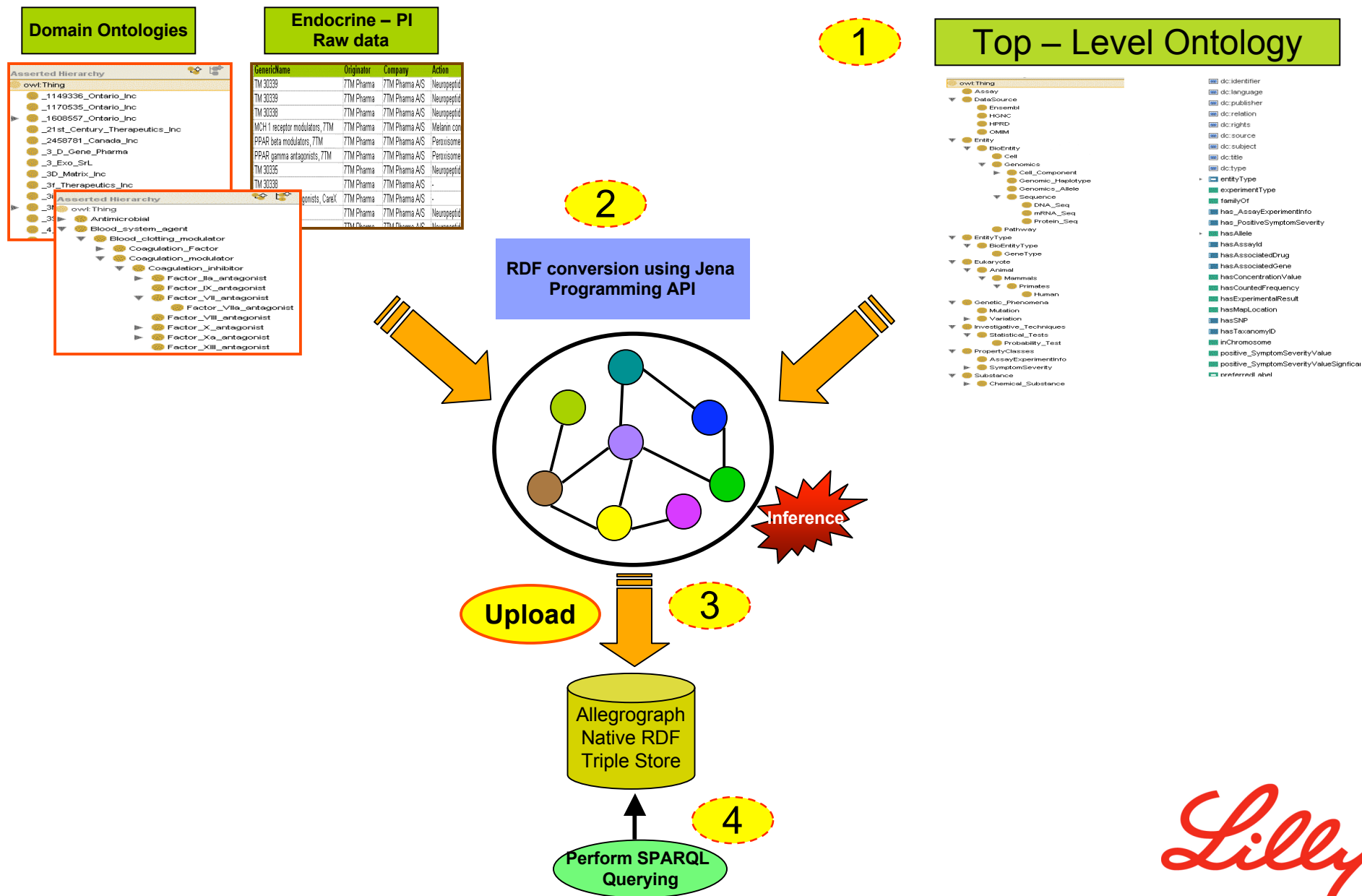
Why Semantic Integration and Knowledge Representation in RDF & OWL?

- Multi Dimensional Data (various data types)
- Synonyms
- Taxonomy Relations
- Easy Inference & Reasoning
- SPARQL – Graph Queries
- Easy addition of new Data types (ex: Anatomy, pathways)
- OWL & RDF inference aware Triple Stores



Workflow

Step2 : RDF Conversion – Loading - Querying



Lilly

Objective 4: Querying - Inference - Reasoning

Will go through Simple to Complex Query examples to test

- **Semantic Integration**
- **Types Of Inference**
- **Drawing contrast to queries and comparing the results without Semantic Integration and Inference**

Lilly

Is Synonym based Inference and Semantic Integration important for Information Retrieval and knowledge mining?

Simple Query

Given Company Name: Applied Molecular Genetics Inc Get MOA's that this company is working in obesity

```
PREFIX TLO: <http://www.lilly.com/POC/TopLevel-ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
Select Distinct ? Endo_MOA ?Company_All_Labels
Where{
    ?Company_Res      TLO:SynonymousLabels      "Applied Molecular Genetics Inc"^^xsd:string .
    ?Company_Res      TLO:preferredLabel         ?Company_Prer_Label .
    ?Company_Res      TLO:SynonymousLabels      ?Company_All_Labels .
    ?Drug_Info        TLO:hasAssociatedCompany  ?Company_All_Labels .
    ?Drug_Info        TLO:hasMOA                ?Endo_MOA
}
```

Case1 : *'Without'* Semantic Integration and Inference

'0' Results

Case2 : *'With'* Semantic Integration and Inference

Amgen	Leptin stimulator
Amgen Inc	Agouti related protein inhibitor
Amgen	Neuropeptide Y antagonist
Amgen Inc	Melanocortin MC4 antagonist

18 Results

Lilly

Is Parent - Child based Inference and Semantic Integration important for Information Retrieval and knowledge mining?

Moderately Complex Query

Given MOA: Neurotransmitter modulator Get Companies that work on this MOA

```

PREFIX TLO: <http:// www.lilly.com/POC/TopLevel -ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX MOA: <http:// www.lilly.com/LifeSciences/ThomsonPharmaMOA #>
Select Distinct ?MOA_labels_preferred ?Endo_CompanyName
Where{
    {
        ?MOA_parent TLO:preferredLabel "Neurotransmitter modulator"^^xsd:string
        ?MOA_parent TLO:preferredLabel ?MOA_labels_preferred .
        ?MOA_parent TLO:preferredLabel ?MOA_labels .
        ?Endo_drugInfo TLO:hasMOA ?MOA_labels .
        ?Endo_drugInfo TLO:hasFinalCompany ?Endo_CompanyName .
    }
    UNION
    {
        ?MOA_parent TLO:SynonymousLabels "Neurotransmitter modulator"^^xsd:string
        ?MOA_parent TLO:hasSubClass ?MOA_child .
        ?MOA_child TLO:SynonymousLabels ?MOA_labels_preferred .
        ?MOA_child TLO:preferredLabel ?MOA_labels .
        ?Endo_drugInfo TLO:hasMOA ?MOA_labels .
        ?Endo_drugInfo TLO:hasAssociatedCompany ?Endo_CompanyName .
    }
}
    
```

Case1 : *'Without'* Semantic Integration and Inference

NeuroSearch A/S	Neurotransmitter modulator
-----------------	----------------------------

1 Result

Case2 : *'With'* Semantic Integration and Inference

NeuroSearch A/S	Neurotransmitter modulator
Glaxo Wellcome plc	Nicotinic ACh receptor antagonist
Merck & Co Inc	Cannabinoid CB1 receptor agonist
Johnson & Johnson	Glutamate receptor antagonist
Bristol-Myers Squibb Co	Neuropeptide Y modulator

Parent – Child Relation

225 Results where a Total of 84 Companies working on 74 MOA's that are subclass of "Neurotransmitter modulator"

SPARQL Query that includes inference (Synonym + Subsumption) combining with Semantic Integration for Information Retrieval and knowledge mining?

Kind of Complex Query

Get All GPCR Modulator (MOA's) that Pfizer is working on including the Compounds , Therapeutic areas and their Phases

```

PREFIX TLO: <http:// www.lilly.com/POC/TopLevel -ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX MOA: <http:// www.lilly.com/LifeSciences/ThomsonPharmaMOA #>
Select Distinct ? Company_preferredLabel ?MOA_labels_preferred ?status ? drugName
where{
    {
        ?Parent_comp TLO:Synonym ousLabels "Pfizer Limited"^^xsd:string .
        ?Parent_comp TLO:Synonym ousLabels ?Parent_comp_Labels .
        ?MOA_parent TLO:Synonym ousLabels "GPCR modulator"^^xsd:string .
        ?MOA_parent TLO:hasSubClass ?MOA_child .
        ?MOA_child TLO:Synonym ousLabels ?MOA_child_labels .
        ?Endo_drugInfo ?Parent_comp_Labels .
        ?Endo_drugInfo ?MOA_child_labels .
        ?Endo_drugInfo ?status .
        ?Endo_drugInfo ?drugName .
        ?MOA_child ?MOA_labels_preferred .
        ?Parent_comp ?Company_preferredLabel .
    }
    UNION
    {
        ?Parent_comp TLO:Synonym ousLabels "Pfizer Limited"^^xsd:string .
        ?Parent_comp TLO:hasSubsidiaryCompany ?Subsidiary_Company .
        ?Subsidiary_Company TLO:Synonym ousLabels ?Subsidiary_Company_Labels .
        ?MOA_parent TLO:Synonym ousLabels "GPCR modulator"^^xsd:string .
        ?MOA_parent TLO:hasSubClass ?MOA_child .
        ?MOA_child TLO:Synonym ousLabels ?MOA_child_labels .
        ?Endo_drugInfo ?Subsidiary_Company_Labels .
        ?Endo_drugInfo ?MOA_child_labels .
        ?Endo_drugInfo ?status .
        ?Endo_drugInfo ?drugName .
        ?MOA_child ?MOA_labels_preferred .
        ?Subsidiary_Company ?Company_preferredLabel .
    }
}

```

Earlier Complex Query Investigation

```
PREFIX TLO: <http://www.illy.com/POC/TopLevel-ontology/#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX MOA: <http://www.illy.com/LifeSciences/ThomsonPharmaMOA#>
Select Distinct ?Company_preferredLabel ?MOA_labels_preferred ?status ?drugName
where{
```

```
{
  ?Parent_comp TLO:SynonymousLabels
  ?Parent_comp TLO:SynonymousLabels
  ?MOA_parent TLO:SynonymousLabels
  ?MOA_parent TLO:hasSubClass
  ?MOA_child TLO:SynonymousLabels
  ?Endo_drugInfo TLO:hasOriginatingCompany
  ?Endo_drugInfo TLO:hasMOA
  ?Endo_drugInfo TLO:hasStatus
  ?Endo_drugInfo TLO:hasDrug
  ?MOA_child TLO:preferredLabel
  ?Parent_comp TLO:preferredLabel
}
```

UNION

```
{
  ?Parent_comp TLO:SynonymousLabels
  ?Parent_comp TLO:hasSubsidiaryCompany
  ?Subsidiary_Company TLO:SynonymousLabels
  ?MOA_parent TLO:SynonymousLabels
  ?MOA_parent TLO:hasSubClass
  ?MOA_child TLO:SynonymousLabels
  ?Endo_drugInfo TLO:hasOriginatingCompany
  ?Endo_drugInfo TLO:hasMOA
  ?Endo_drugInfo TLO:hasStatus
  ?Endo_drugInfo TLO:hasDrug
  ?MOA_child TLO:preferredLabel
  ?Subsidiary_Company TLO:preferredLabel
}
```

```
"Pfizer"^^xsd:string .
?Parent_comp_Labels .
"GPCR modulator"^^xsd:string .
?MOA_child .
?MOA_child_Labels .
?Parent_comp_Labels .
?MOA_child_Labels .
?status .
?drugName .
?MOA_labels_preferred .
?Company_preferredLabel .
```

Concerns with term "GPCR Modulator"

- "GPCR modulator", the exact term is not present in the Endocrine Competitive Intelligence data, so synonyms of GPCR modulator must be considered
- Pfizer might be working on subclasses of "GPCR Modulator"
- "GPCR Modulator" subclasses might be having semantic and syntactic variations

Concerns with term "Pfizer Limited"

- "Pfizer Limited", the exact term is not present in Endocrine – PI raw data
- Sometimes "Pfizer" might not work but its subsidiaries will be working on "GPCR Modulator"
- These subsidiary companies might also work on subclasses of "GPCR Modulators"
- These subsidiary companies might have semantic and syntactic variations

Case1 : 'Without' Semantic Integration and Inference

0 Results

Case2 : 'With' Semantic Integration and Inference

Company Name	MOA	Status	Drug Name	Therapeutic Area
Pfizer Inc	Adrenoceptor agonist	Phase I Clinical	CP 331684	Obesity
Alanex Corp	Neuropeptide Y receptor antagonist	Discontinued	neuropeptide Y antagonists A	Obesity
Parke-Davis & Co	CCK receptor antagonist	No Development Reported	PD 145942	Obesity
Parke-Davis & Co	Neuropeptide Y1 receptor antagonist	No Development Reported	PD 160170	Obesity

26 Results

Tools Used in Project Execution

Setting up the programming & storage environment

Programming Environment



Semantic Web API



Triple Store



Text Mining

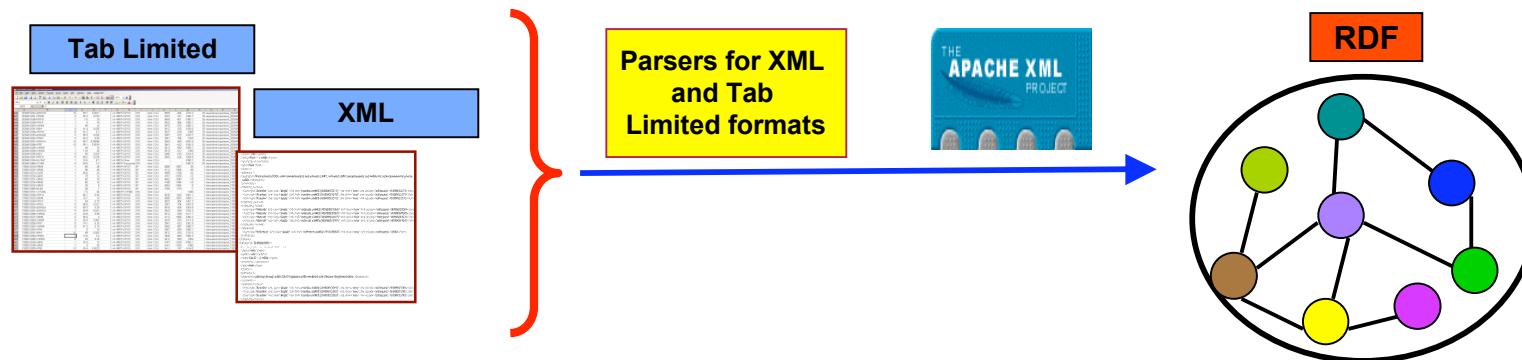


The Specialist
Lexical
Tools

Ontology Development

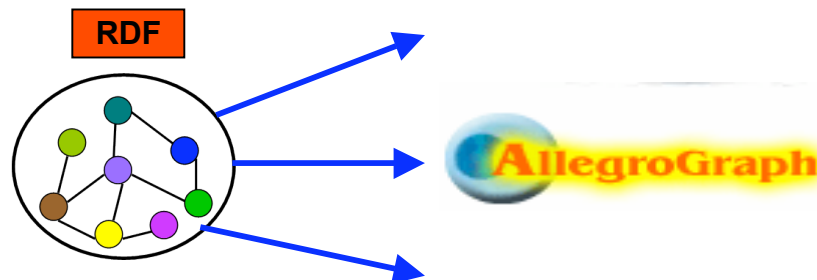


Writing parsers to convert data into triples



Loading triples to stores

Custom loaders using API
specific to triple store



What are the key lessons learned with this Project?

Key Lessons

- Thomson – Pharma's Company & MOA Ontologies are capable for semantic integration and reasoning
- Semantic Integration is achieved at two levels :
 - Raw Competitive Intelligence data can be mapped (Instance mapping) to Ontologies using NLP techniques
 - Efficient semantic integration by using RDF and OWL
- Powerful complex data modeling achieved by using graph principles inherent in RDF
- Ontology development tools (ex: Protégé) can be used for manual ontology enrichment
- Knowledge Discovery is possible with inference and reasoning on Competitive Intelligence data
- Triple stores (ex: Allegrograph, Oracle 11g) have sufficient inference engines for reasoning
- Easy translation of questions to graph queries using SPARQL

Risks identified in this project

- Finding mapping methods (NLP) from 'Raw Intelligence' to Ontologies

Ex:

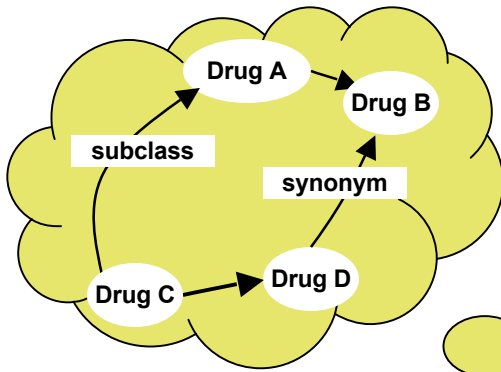
Instance Data	Term In Ontology
Melanocortin-4 receptor antagonists	MC4R antagonist
5 Hydroxytryptamine 1B agonist	5-HT1b agonist
Serotonin 2A receptor antagonists	5-HT2a antagonist

- Noise inherent in Thomson – Pharma XML files

The Lilly logo, featuring the word "Lilly" in a red, cursive script font.

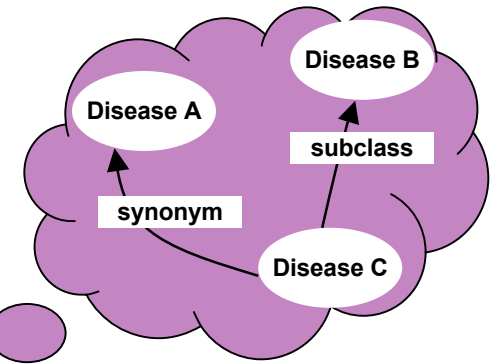
Further Improvement to Competitive Intelligence

Drug Ontology

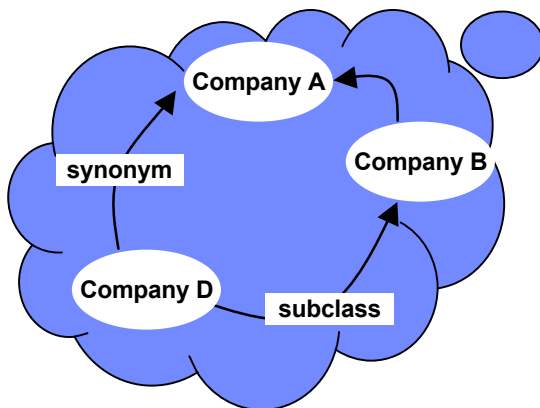


A Massive Competitive Intelligence RDF Information Space can be created and used for knowledge discovery and mining

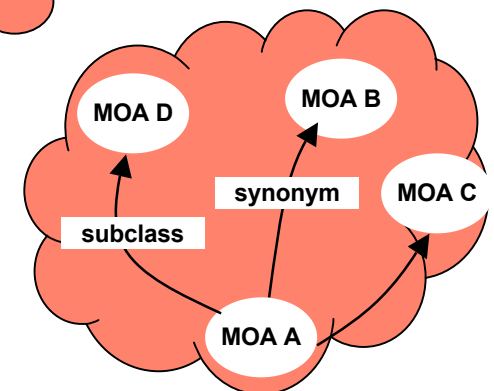
Disease Ontology



Company Ontology



MOA Ontology



Overall Conclusion of Project

- Semantic Integration (instance mapping using NLP) coupled with RDF data model was successful in answering questions in Competitive Intelligence
- Ontologies provide a powerful framework in providing dictionaries and taxonomical relations that help to reason and inference the data for knowledge discovery
- Manual curation is a tedious, error prone and labor intensive-task
 - A semi-automated intelligent computer-based solution that utilizes Ontologies, Semantic Integration and NLP could drastically reduce manual curation process and maintain high quality information

Final Thoughts

- SPARQL is an essential component for accessing/constraining SW data, but by itself it is insufficient for the discovery of new associations and mechanisms in whole biological systems. Many of the major challenges within pharma R&D are precisely of this type!
- Instance based Semantic integration
- Graph Theoretic principles (Ex: Clustering, Page-rank Algorithms)
- Applying statistics

Contributors

Project Members

Ranga Chandra Gudivada

Jacob Koehler

Joseph F Ferrara

William Sanchez

Business Customers

Yiqun Helen Li

Yuhao Lin

Managers

Vaibhav A Narayan

Christopher Otto

Susie Stephens



Manuscript Under Preparation

Title : Semantic Integration of the Competitive Drug Landscape

Target Journal: Journal of Biomedical Informatics

Lilly

Thanks

Any Questions?

Lilly