

The Integration of Biological Data Using Semantic Web Technologies

Susie Stephens
Principal Product Manager, Life Sciences
Oracle

susie.stephens@oracle.com

Outline

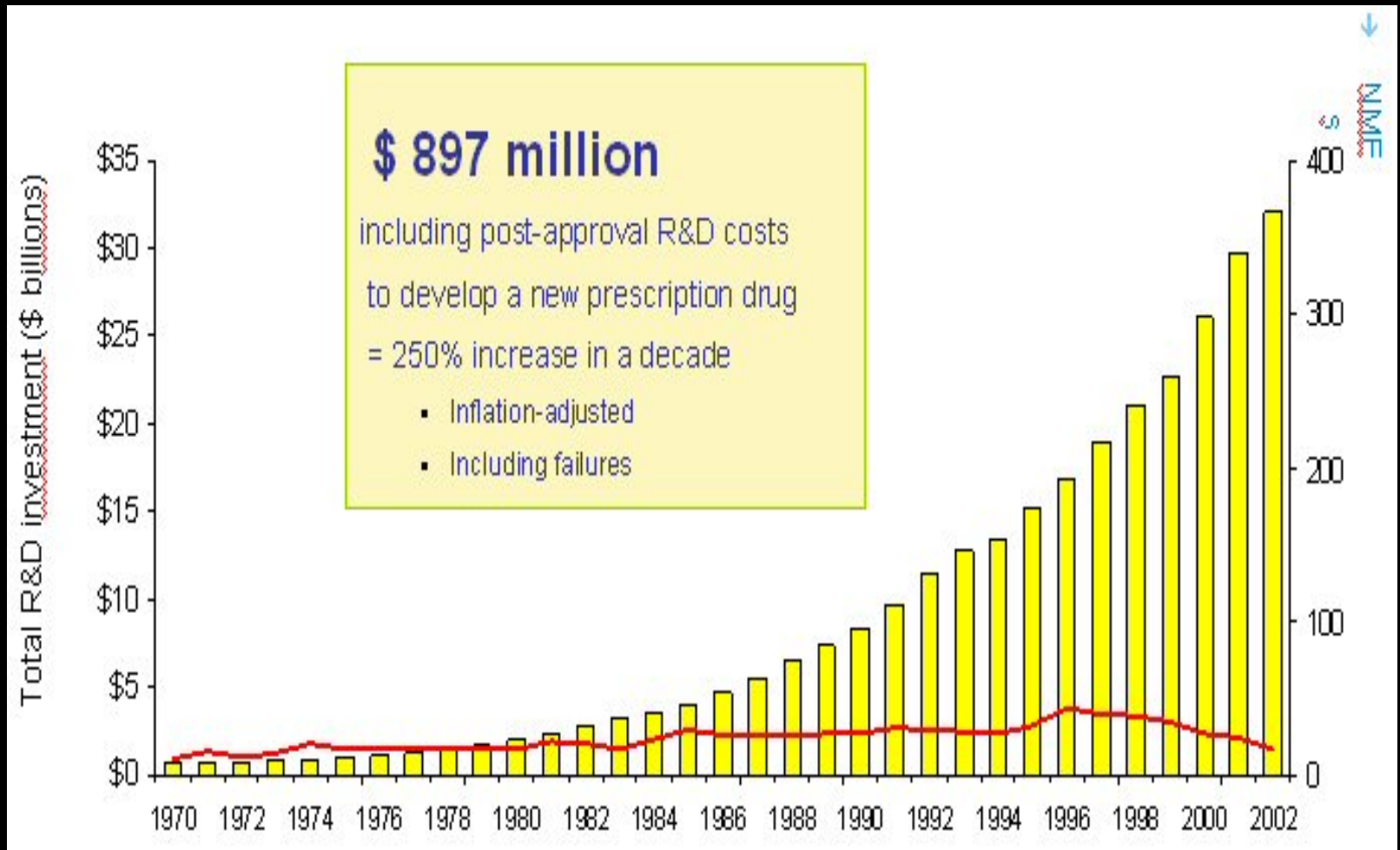
- Complexity of Biological Data
- Oracle's RDF Data Model
- Life Sciences Use Cases

The Complexity of Biological Data

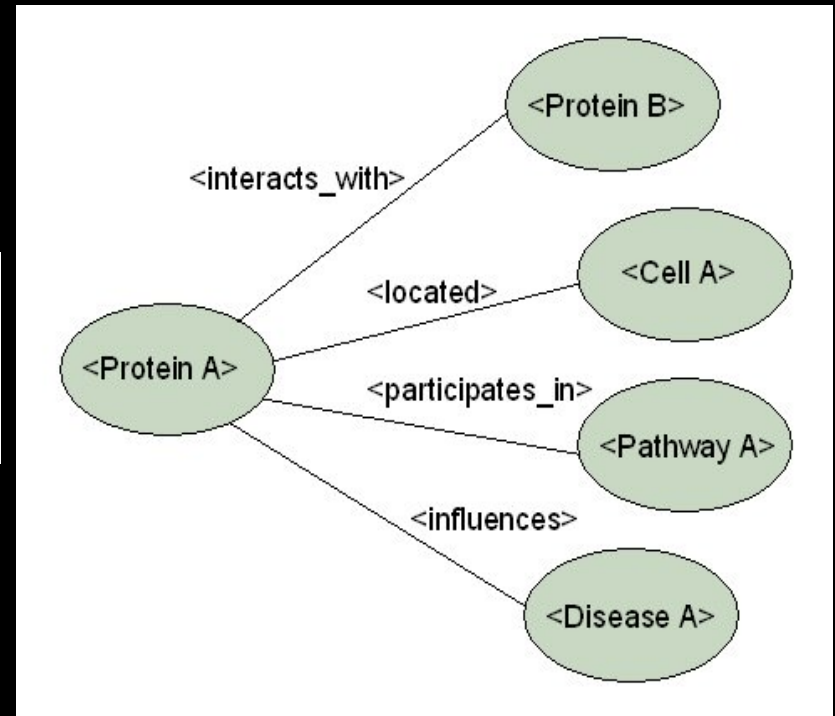
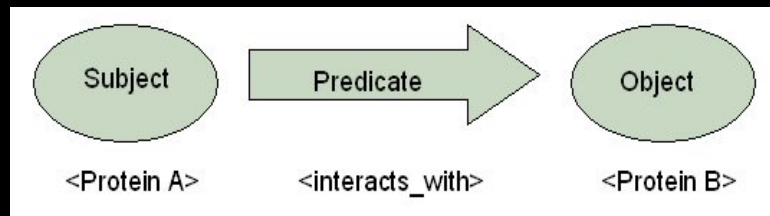
The image displays a collage of biological data visualization tools and databases:

- Cr3D 3.0:** A 3D molecular model of a protein structure, showing a complex fold with various colored domains.
- ENZYME: EC 3.1.1.7:** A detailed entry for Acetylcholinesterase (EC 3.1.1.7) from the Enzyme Commission. It includes alternative titles, a description, and references. The text states: "Acetylcholine + H₂O = choline + acetate. -!- Acts on a variety of acetic esters. -!- Also catalyzes transesterifications." It also lists databases like BRENDA, EMB/EMBL, and KEGG.
- Metabolic Pathway:** A complex diagram titled "Glycerolipid metabolism - Reference pathway" showing the conversion of various lipids and phospholipids into products like choline, ethanolamine, and glycerol. Key enzymes and cofactors like CDP-ethanolamine and CDP-diacylglycerol are shown.
- Gene Database:** A screenshot of a database entry for the gene **ACETYLCHOLINESTERASE** (ACHE). It provides the gene map locus **7q22** and references to scientific literature, such as Coates and Simpson (1972) and Rotondo et al (1988).
- Sequence Alignment:** A screenshot of a sequence alignment viewer showing multiple protein sequences aligned against a reference sequence. The alignment is visualized with colored bars above the sequence lines.
- 3D Protein Structure:** A 3D visualization of a protein structure, likely related to the enzyme discussed in the other panels, showing its spatial arrangement and potential active sites.

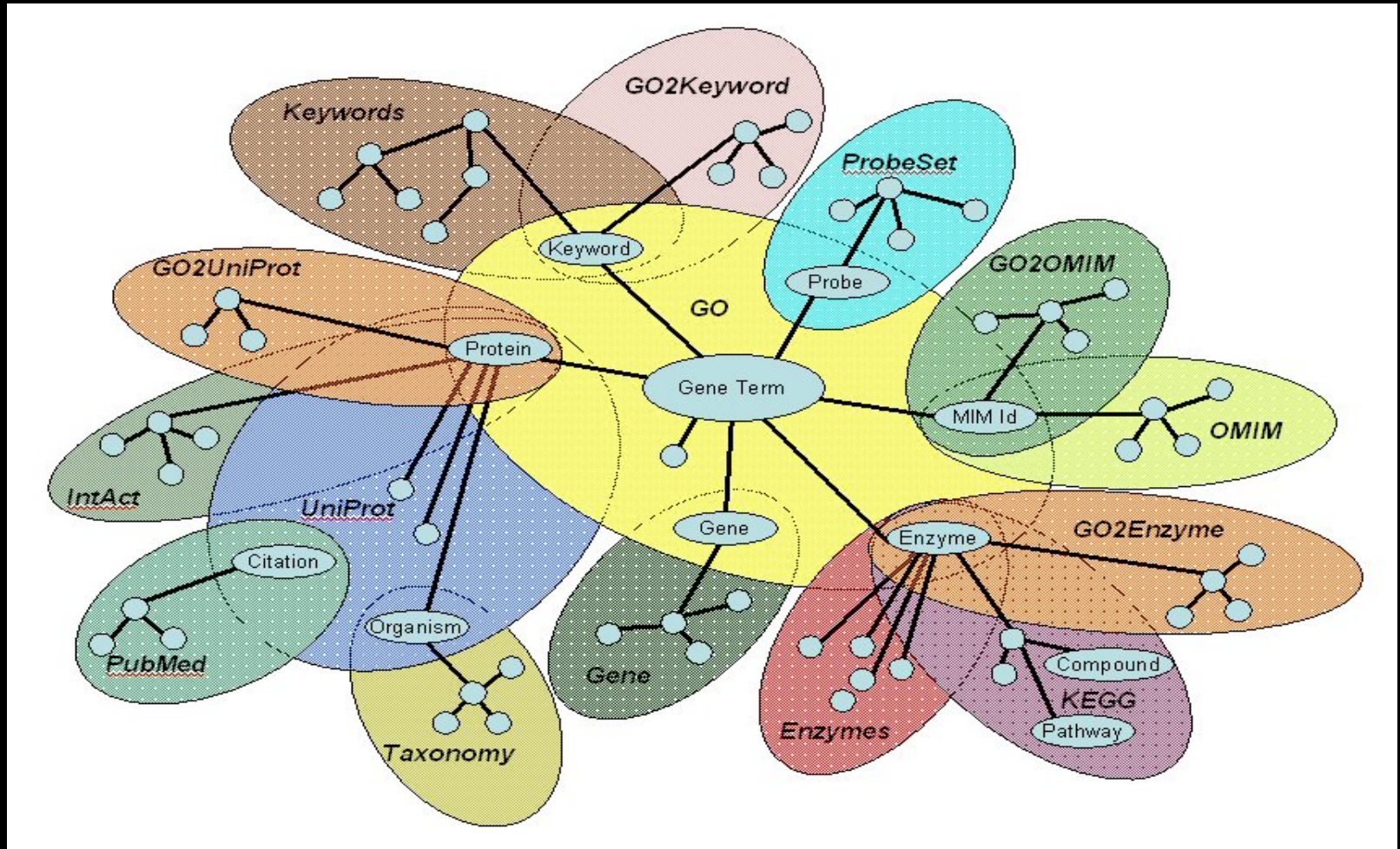
Pharmaceutical Productivity



RDF Triples in Life Sciences



The Semantic Web Vision



Outline

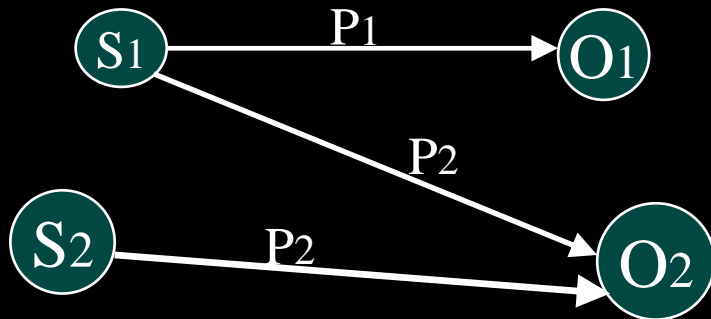
- Life Sciences Data
- Oracle's RDF Data Model
- Use Cases

Oracle and RDF: Motivation

- Customer requests
- RDF (and OWL) are maturing
- Oracle supports open standards
- Complements Oracle's information management approaches
- Ability to leverage existing technologies

Oracle RDF Data Model

- Support for RDF and RDFS
- Object-relational implementation
- Subjects and objects are re-used
- Links represent complete RDF triples



RDF Triples:

- {S1, P1, O1}
- {S1, P2, O2}
- {S2, P2, O2}

SPARQL-like Query Capability

- A table function allows a graph query to be embedded in a SQL query
- Searches for an arbitrary pattern against the RDF data
- Includes inferencing based on RDF, RDFS, and user-defined rules

Enterprise Functionality

- Real Application Clusters (RAC), Security
- Multi-threaded, parallel processing, indexed, etc.
- Performance testing with UniProt

	Q1	Q2	Q3	Q4	Q5	Q6
10 M Triples	0.86	< 0.01	< 0.01	0.03	0.18	0.46
20 M Triples	0.95	< 0.01	< 0.01	0.03	0.19	0.47
40 M Triples	0.96	< 0.01	< 0.01	0.03	0.18	0.47
80 M Triples	1.03	< 0.01	< 0.01	0.03	0.20	0.49
Maximum σ	.054	0.002	0.002	.011	.065	0.07

Units in seconds

Image Search

“Find me all DICOM images that contain the term ‘Jaw’”

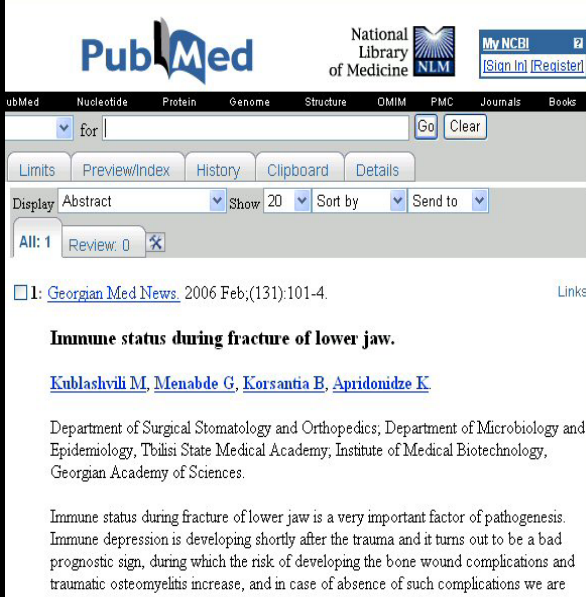
- Map relationships to terms using RDF triples
 - ‘Mandible’, sameAs’, ‘Jaw’
 - ‘Maxilla’, ‘partOf’, ‘Jaw’



Text Search

“Find me all papers that contain the term ‘Jaw’”

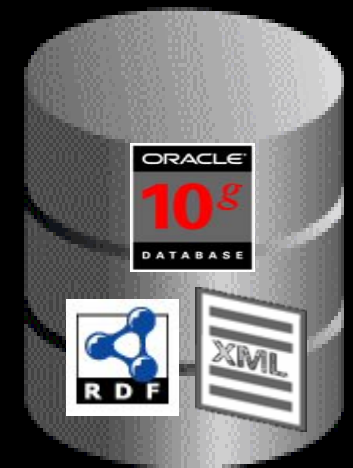
- Map relationships to terms using RDF triples
 - ‘Mandible’, sameAs’, ‘Jaw’
 - ‘Maxilla’, ‘partOf’, ‘Jaw’



The screenshot shows the PubMed website interface. At the top, there is the PubMed logo and the National Library of Medicine (NLM) logo. Below the search bar, there are navigation tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The search results are displayed in a table format. The first result is from 'Georgian Med News' (2006 Feb;(131):101-4) with the title 'Immune status during fracture of lower jaw'. The authors listed are 'Kublashvili M, Menabde G, Korsantia B, Apridonidze K'. The abstract text is partially visible, starting with 'Immune status during fracture of lower jaw is a very important factor of pathogenesis. Immune depression is developing shortly after the trauma and it turns out to be a bad prognostic sign, during which the risk of developing the bone wound complications and traumatic osteomyelitis increase, and in case of absence of such complications we are found with significant retention of a bone fracture of lower jaw. We have consid...

Data Integration

- SQL / RDBMS
 - Concise, efficient transactions
 - Transaction metadata is embedded or implicit in the application or database schema
- XQuery / XML
 - Transaction across organizational boundaries
 - XML wraps the metadata about the transaction around the data
- SPARQL / RDF
 - Information sharing with ultimate flexibility
 - Enables semantics as well as syntax to be embedded in documents



Download the Database!

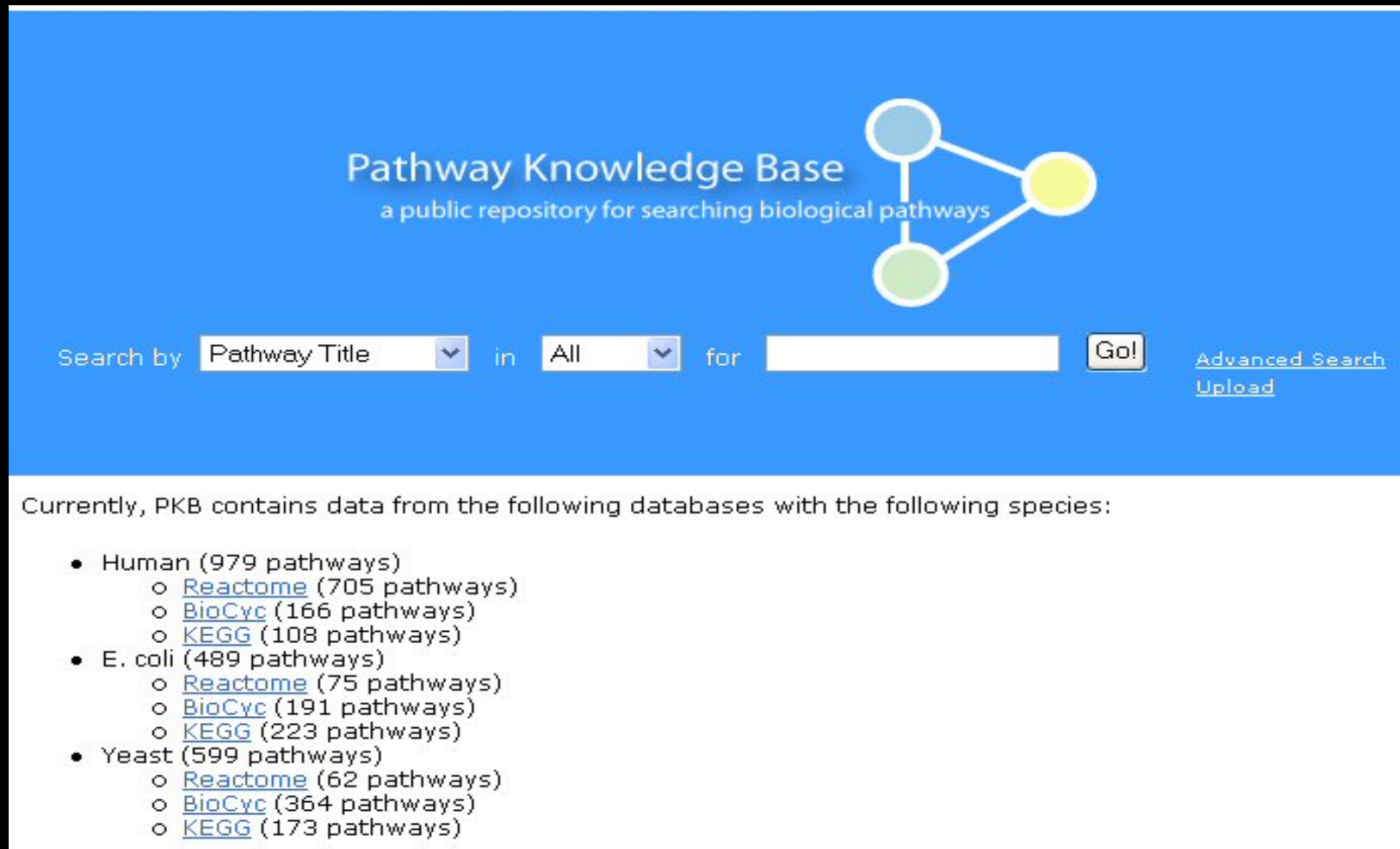
Oracle Database Enterprise Edition 10g Release 2

<http://www.oracle.com/technology/software/products/database/oracle10g/index.html>

Outline

- Life Sciences Data
- Oracle's RDF Data Model
- Use Cases

Stanford University Use Case



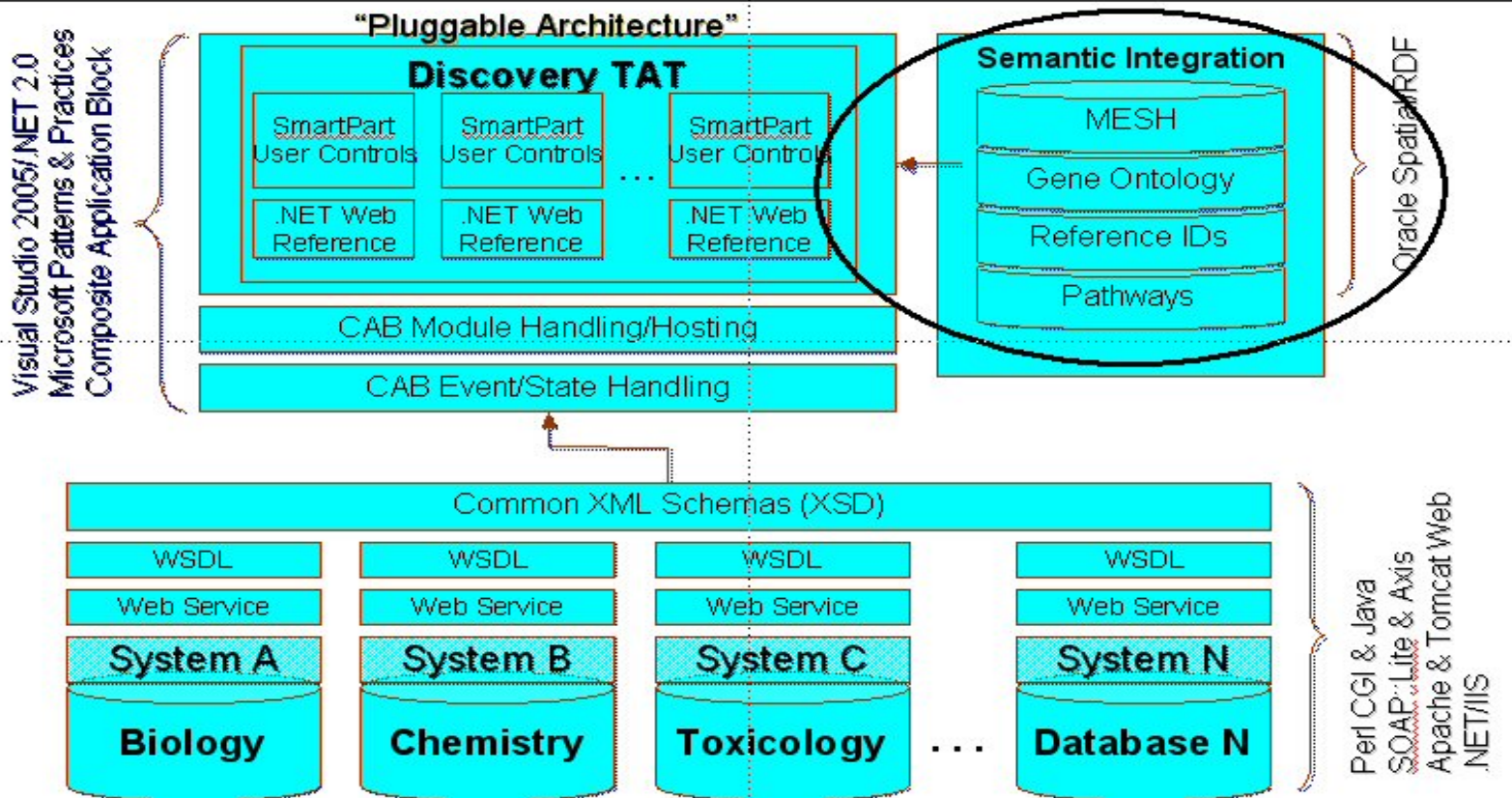
The screenshot shows the Pathway Knowledge Base (PKB) website. The header features the text "Pathway Knowledge Base" and "a public repository for searching biological pathways" next to a logo of three interconnected nodes (blue, green, and yellow). Below the header is a search bar with the following elements: "Search by" followed by a dropdown menu set to "Pathway Title", "in" followed by a dropdown menu set to "All", "for" followed by an empty text input field, and a "Go!" button. To the right of the search bar are links for "Advanced Search" and "Upload".

Currently, PKB contains data from the following databases with the following species:

- Human (979 pathways)
 - [Reactome](#) (705 pathways)
 - [BioCyc](#) (166 pathways)
 - [KEGG](#) (108 pathways)
- E. coli (489 pathways)
 - [Reactome](#) (75 pathways)
 - [BioCyc](#) (191 pathways)
 - [KEGG](#) (223 pathways)
- Yeast (599 pathways)
 - [Reactome](#) (62 pathways)
 - [BioCyc](#) (364 pathways)
 - [KEGG](#) (173 pathways)

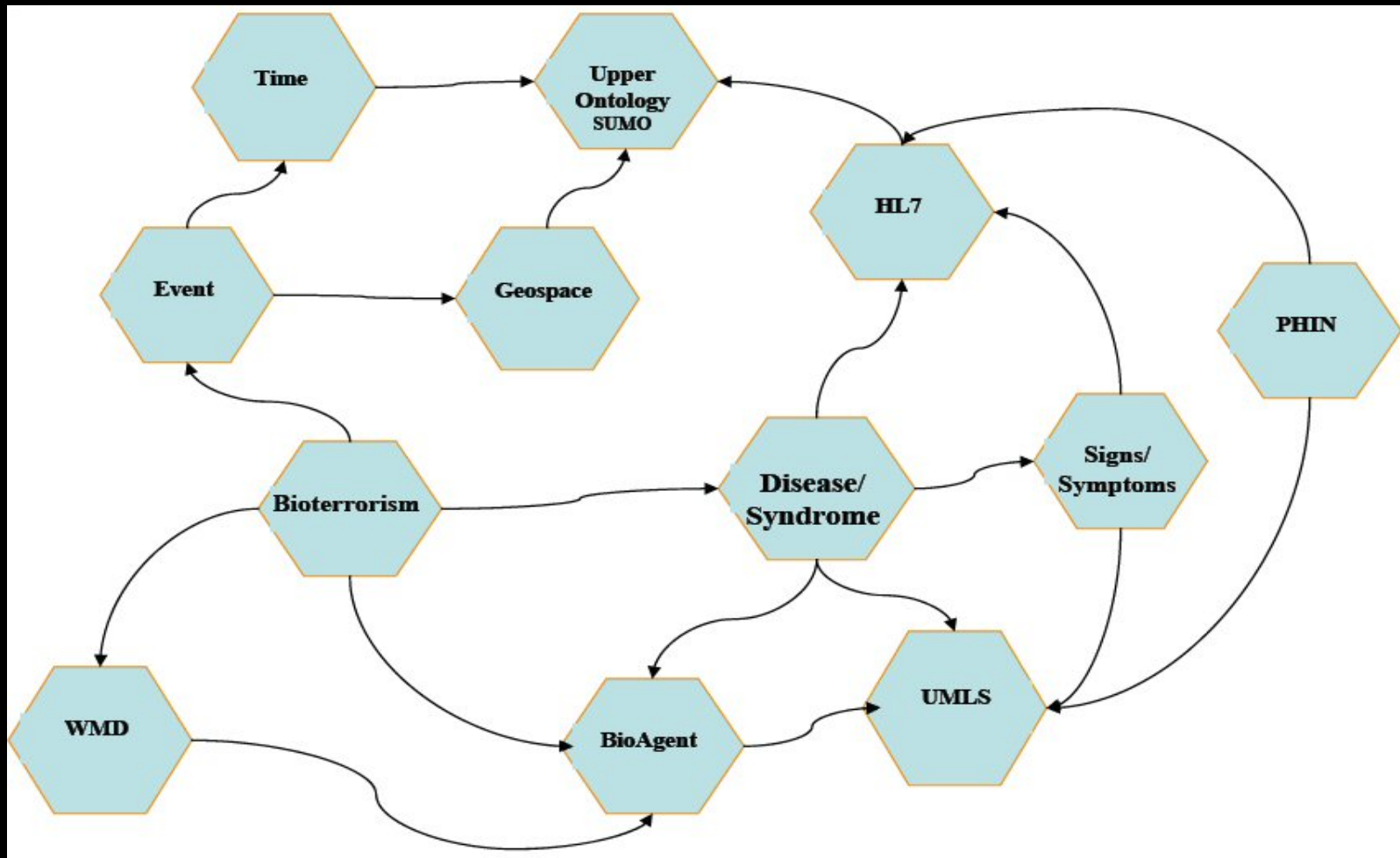
Eli Lilly Use Case

Discovery TAT Service-Oriented Architecture



ORACLE

University of Texas Health Science Center Use Case



BioRDF

ESW Wiki [Login](#)

[HCLSIG BioRDF Subgroup](#) [Tasks](#)

[FrontPage](#) [RecentChanges](#) [FindPage](#) [HelpContents](#) [Tasks](#)

[Edit \(Text\)](#) [Edit \(GUI\)](#) [Info](#) [Attachments](#) [More Actions:](#)

Active Tasks

- [/Reagents \(Status\)](#)
- [/SenseLab](#)
- [/Using SW Technologies to Find Small Molecules that Bind to Proteins](#)
- [/Gene Neural related gene data](#)
- [/OMIM Neural diseases](#)
- [/Natural Language Processing and RDF](#)
- [/Ligand-Receptor Interaction, Molecular Interaction Networks, Ontology Evolution](#)
- [/Vocabulary Requirements](#)

Proposed Tasks

- [/Brain Connectivity](#)
- [/Brain Atlas condition to scans](#)
- [/Protein Neural related protein data](#)
- [/Ruby On Rails and ActiveRDF](#)

Summary

- The Semantic Web provides the ability to more easily integrate heterogeneous data
- Oracle has a scalable, secure, highly-available RDF Data Model
- Adoption of Semantic Web technologies is accelerating
- Make your data sharable, make it available in RDF