

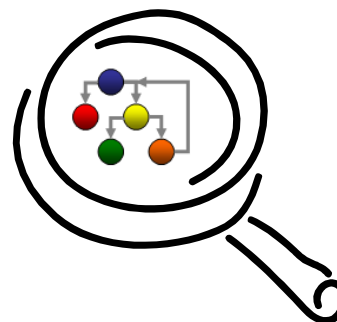
SLabSearch™

Drug Discovery (Re)Search Tool

Carlos S. Zamudio
Semantic Laboratories

www.semanticlaboratories.com

April 2006



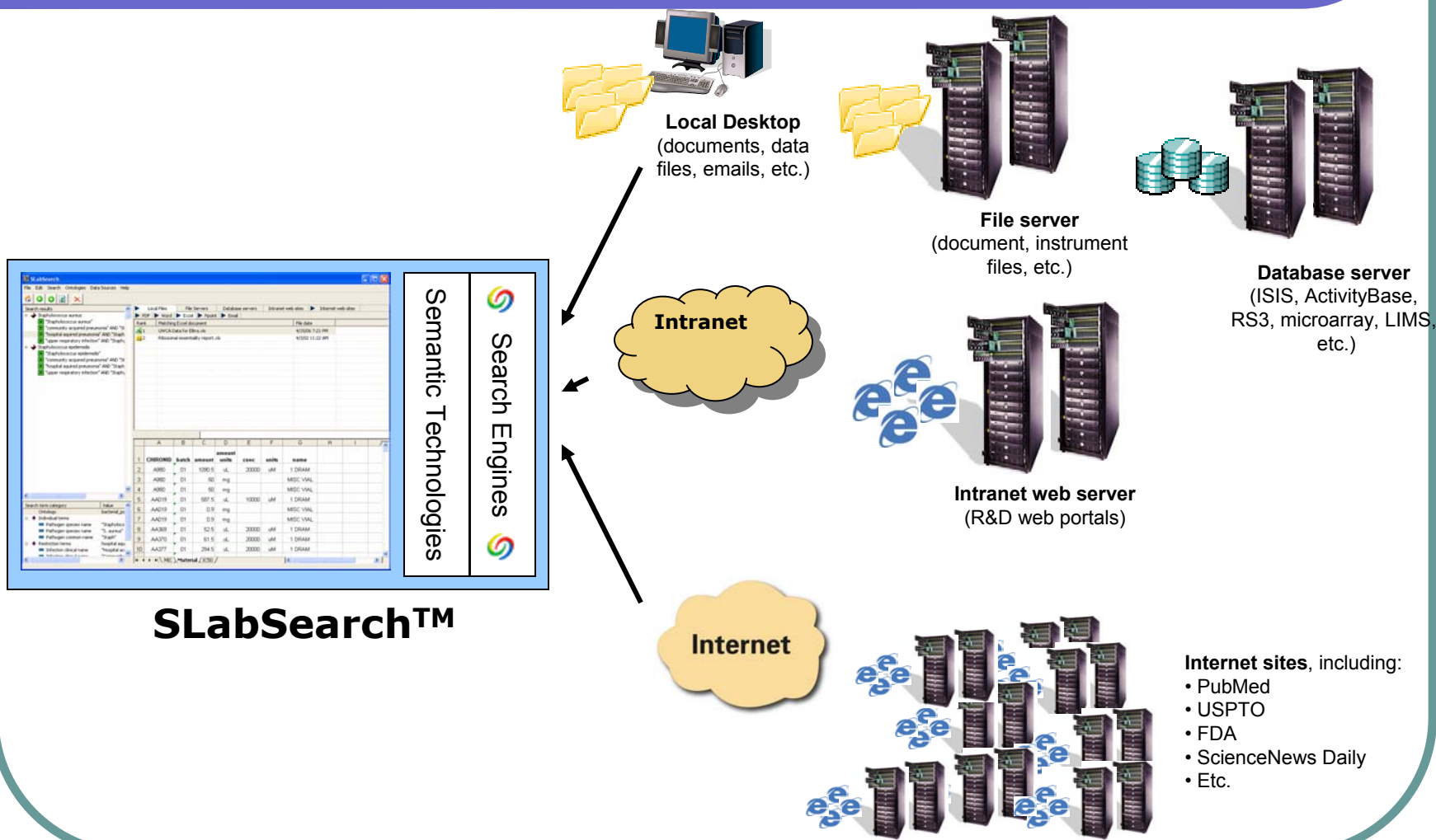
Major Drug Discovery R&D Activities

- Data and knowledge capture
- Data and knowledge integration and analysis
- Data and knowledge dissemination
- Data and knowledge re-use
- Decision making based on currently available information

Significant barriers exist to maximizing the potential of these integrated activities



SLabSearch™ – Semantic Search Utility



Barriers to Data and Knowledge Re-use and Integration

- Determining *whether* the data or knowledge exists (inside or outside the organization)
- Determining *where* the data or knowledge resides
- Determining *which* query tool to use (and the query syntax) to extract them from the repository
- Determining the *exact terminology* (and spelling) used by the authors to label the data or knowledge
- Finding the appropriate tool for *retrieving* data for review and evaluation
- Converting data formats for *integrating* data with other data and knowledge
- Identifying the appropriate repository for *publishing* the new synthesized data and knowledge for further integration

These barriers impact research productivity, cost and limit the ability to leverage organizational knowledge



The Research Cost of These Barriers

- Decreased data and knowledge re-use across the organization
- Less informed decision making by project teams
- Decreased productivity
 - Scientists are required to master the structure of the data management solutions to form correct queries
 - Specialized informatics support staff are required for interfacing distributed data sources
 - Perception that it is sometimes quicker to re-do the experiment

Access to relevant and timely information affects the productivity and quality of decision making by research project teams which determines the success of programs



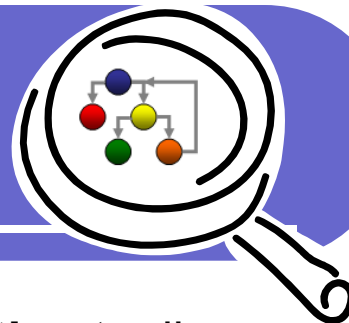
Common Data Integration IT Solutions

- There are a variety of approaches that have been implemented to integrate distributed research data and provide centralized access:
 - Ad-hoc (i.e., nothing formal)
 - Create shared file servers with common folder hierarchies
 - Require that (all) data be published in relational databases
 - Create data interoperability through:
 - Federation of distributed databases with a single query interface
 - Centralization of data into a single database (or data warehouse) with a single query interface
 - Other data transformation approaches which tend to be (very) high maintenance activities and often require that data be transformed into multiple formats for integration.

We need a way to allow data to reside in its native format containers, yet be able to assess its relevance and easily locate it for integration...SLabSearch™

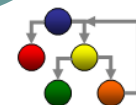


SLabSearch™



- SLabSearch™ is a research enterprise search application to discover, retrieve and integrate research data and knowledge across the R&D enterprise using data semantics and search engine technologies
- SLabSearch™ can accommodate data that is **widely distributed** and within their **native formats** (files, relational databases, web sites, etc.)
- SLabSearch™ can search data that **resides** on local hard drives, shared file servers, relational databases, intranet web sites, document management systems and relevant public web sites (e.g., PubMed, FDA, USPTO) while maintaining **security**
- SLabSearch™ provides wizard-style desktop user interfaces to discover complex drug discovery data using **conceptual queries** (using semantic search ontologies) that synthesize the search terms and term relationships reflected in complex scientific concepts

The search paradigm popularized by Google has revolutionized how we look at information acquisition and integration...



Search Interface

Tabs for each data source Collection searched

List of query expressions used in search

Search results for each Collection searched

Semantics of search terms used in query

Contents of matching document (e.g., PDF)

The screenshot displays the SlabSearch application window. At the top, there are menu options: File, Edit, Search, Ontologies, Data Sources, and Help. Below the menu is a toolbar with icons for home, refresh, search, and close. The main interface is divided into several sections:

- Search results:** A tree view on the left shows search results for "Staphylococcus aureus" and "Staphylococcus epidermidis". Each entry is accompanied by a list of query expressions, such as "community acquired pneumonia" AND "Staphylococcus aureus".
- Ranking Table:** A table in the center lists search results with columns for Rank, Matching PDF document, and File date. The results are numbered 1 through 25.
- Document Preview:** A large window at the bottom displays the content of a matching document, which is a page from the "JOURNAL OF combinatorial CHEMISTRY". The page includes the journal title, volume and issue information, and the title of a review article: "Comprehensive Survey of Combinatorial Library Synthesis: 2003".
- Search Term Category:** A table at the bottom left lists search term categories and their values. For example, "Ontology" is "bacterial_patho", "Pathogen species name" is "Staphylococcus", and "Infection clinical name" is "Community-acquired pneumonia".



Search Interface - PubMed

The screenshot shows the SlabSearch application window. The search results are displayed in a table with columns for Rank and Matching PubMed entries. The results list 15 entries, with the first entry being the most relevant to the search terms.

Rank	Matching PubMed entries
1	Hyperinfectibility: An Extremely Sticky Phenotype of Klebsiella pneumoniae Associa...
2	Association between rmpA and magA Genes and Clinical Syndromes Caused by Klebsie...
3	Effect of telithromycin and azithromycin on nasopharyngeal bacterial flora in patients ...
4	Roles of the active site water, histidine-303 and phenylalanine-396 in the catalytic me...
5	Synthesis and evaluation of antimicrobial and anticonvulsant activities of some new 3-...
6	[The pathogens of atypical respiratory infections and asthma]
7	Streptococcus pneumoniae-associated cellulitis in a two-month-old Domestic Shorthair ...
8	The emergency department triage of community-acquired pneumonia project data and...
9	Antibiotic surveillance of Streptococcus pneumoniae in Mississippi.
10	ADP-ribosylating and vacuolating cytotoxin of Mycoplasma pneumoniae represents uni...
11	Azithromycin iv pharmacodynamic parameters predicting Streptococcus pneumoniae kli...
12	A Dominant Complement Fixation Pathway for Pneumococcal Polysaccharides Initiated...
13	[Medical indications and effectiveness of the pneumococcal conjugate vaccine.]
14	[C-reactive protein, leukocyte count and ESR in the assessment of severity of commu...
15	Response to polysaccharide antigens in patients with ataxia-telangiectasia.

PubMed
Collection
search results

Integrated Web
browser for
viewing web
pages



Search Interface - USPTO

The screenshot shows the SlabSearch application window. The search results are displayed in a table with columns for Rank, Matching document, and File date. The results are filtered by Staphylococcus aureus and Staphylococcus epidermidis. The detailed view shows the US Patent & Trademark Office Patent Application Full Text and Image Database. The patent application is for the identification of essential genes in microorganisms, filed by Wang, Liangsu; et al. on February 12, 2004. The abstract describes the use of antisense nucleic acids to identify and develop antibiotics.

Rank	Matching document	File date
1	Specific and universal probes and amplification primers to rapidly d...	2/24/05 12:00 AM
2	Process and composition for inhibiting growth of microorganisms	10/28/04 12:00 AM
3	Methods and reagents for preventing bacteremias	7/29/04 12:00 AM
4	Identification of essential genes in microorganisms	2/12/04 12:00 AM
5	Specific and universal probes and amplification primers to rapidly d...	9/25/03 12:00 AM
6	Administration of negamycin or deoxynegamycin for the treatment...	6/12/03 12:00 AM
7	Identification of essential genes in prokaryotes	5/23/02 12:00 AM
8	SPECIFIC AND UNIVERSAL PROBES AND AMPLIFICATION PRIMER...	5/9/02 12:00 AM

US PATENT & TRADEMARK OFFICE
PATENT APPLICATION FULL TEXT AND IMAGE DATABASE

[Help](#) [Home](#) [Boolean](#) [Manual](#) [Number](#) [PTDLs](#)
[Hit List](#) [Prev](#) [Next](#) [Bottom](#)
[View Shopping Cart](#) [Add to Shopping Cart](#)
[Images](#)

(4 of 8)

United States Patent Application 20040029129
Kind Code A1
Wang, Liangsu; et al. February 12, 2004

Identification of essential genes in microorganisms

Abstract

The sequences of antisense nucleic acids which inhibit the proliferation of prokaryotes are disclosed. Cell-based assays which employ the antisense nucleic acids to identify and develop antibiotics are also disclosed. The antisense nucleic acids can also be used to identify proteins required for proliferation, express these proteins or portions thereof, obtain antibodies capable of specifically binding to the expressed proteins, and to use those expressed proteins as a screen to isolate candidate molecules for rational drug discovery programs. The nucleic acids can also be used to screen for homologous nucleic acids that are required for proliferation in cells other than *Staphylococcus aureus*, *Salmonella typhimurium*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*. The nucleic acids of the present invention can also be used in various assay systems to screen for proliferation required genes in other organisms.

Inventors: Wang, Liangsu, (San Diego, CA); Zamudio, Carlos, (La Jolla, CA); Malone, Cheryl.

US Patent Office
Collection
search results

Integrated Web
browser for
viewing web
pages



Search Interface - Excel

The screenshot displays the SlabSearch application window. On the left, a search results tree shows filters for *Staphylococcus aureus* and *Staphylococcus epidermidis*. The main pane shows a list of matching documents with columns for Rank, Matching document, and File date. Below this, an Excel spreadsheet is integrated, displaying a table with columns for ID, batch, amount, units, conc, units, and name. The table contains 22 rows of data. At the bottom left, a search term category panel shows filters for Individual terms and Restriction terms.

ID	batch	amount	units	conc	units	name
A980	01	1090.5	uL	20000	uM	1 DRAM
A980	01	50	mg			MISC VIAL
A980	01	50	mg			MISC VIAL
AA019	01	587.5	uL	10000	uM	1 DRAM
AA019	01	0.9	mg			MISC VIAL
AA019	01	0.9	mg			MISC VIAL
AA369	01	52.5	uL	20000	uM	1 DRAM
AA370	01	61.5	uL	20000	uM	1 DRAM
AA377	01	294.5	uL	20000	uM	1 DRAM
AAA905	01	286	uL	10000	uM	MINITUBE
AAA905	01	4	mg			MINITUBE
AB124	01	2000	mg			MISC VIAL
AB124	01	573.348	uL	20000	uM	1 DRAM
AB738	01	300.07	uL	20000	uM	1 DRAM
AB738	01	20	mg			MISC VIAL
AB771	01	253.098	uL	20000	uM	1 DRAM
AB771	01	50	mg			MISC VIAL
AC541	01	115.404	uL	20000	uM	1 DRAM
AD540	01	565.693	uL	20000	uM	1 DRAM
AD815	01	315.358	uL	20000	uM	1 DRAM
AD815	01	73.5	mg			1 DRAM

Excel Data Collection search results

Integrated Excel Application for data extraction



Search Interface – Google Desktop

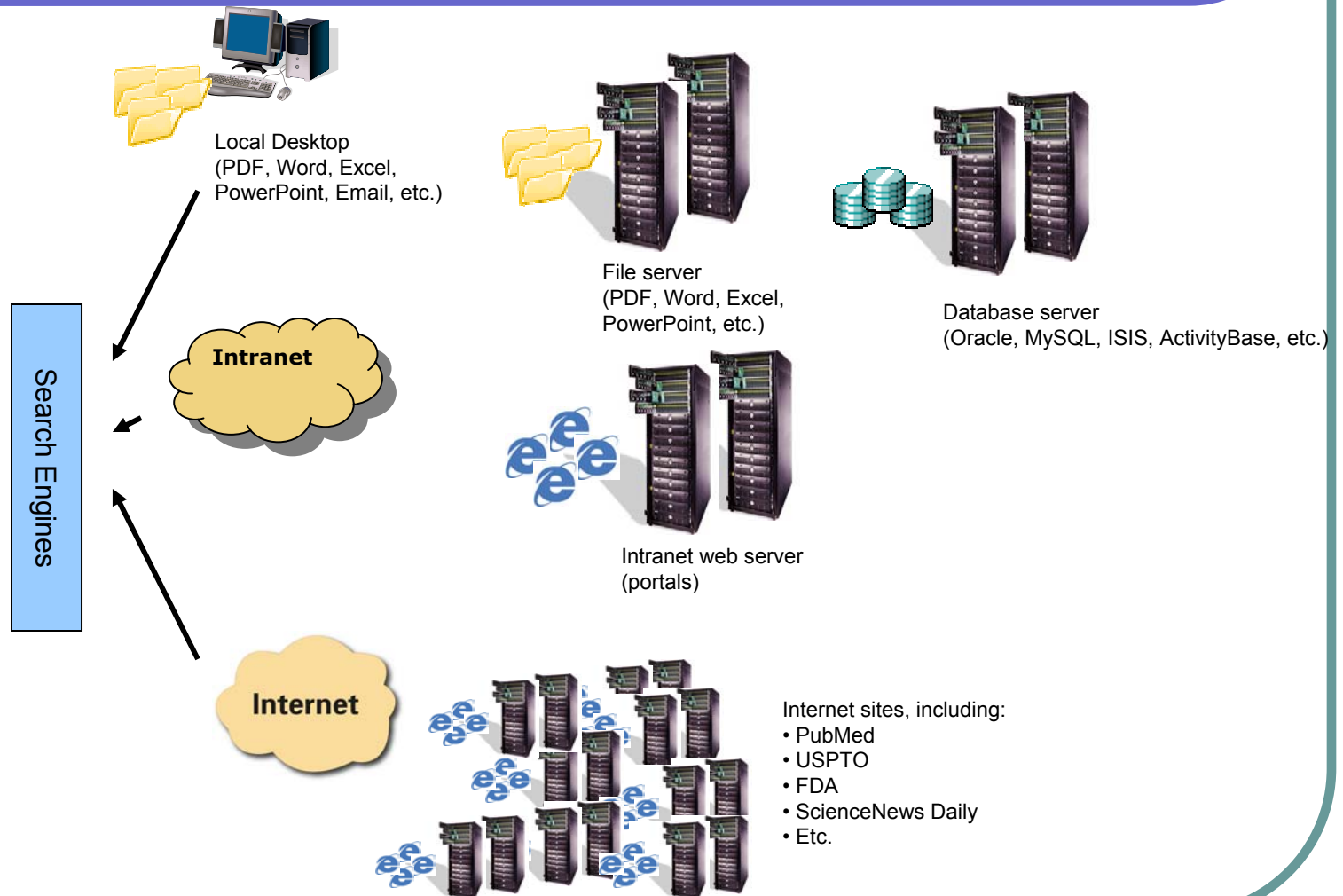
The screenshot displays the SlabSearch application window. The top menu bar includes File, Edit, Search, Ontologies, Data Sources, and Help. The main interface is divided into several sections:

- Search results:** A tree view on the left shows search results for "Staphylococcus aureus" and "Staphylococcus epidermidis".
- Local Files:** A table in the center lists matching email documents with columns for Rank, Matching Email document, and Email date.
- Google Desktop:** An integrated search interface at the bottom shows the Google logo, search bar with the query "Staphylococcus aureus" "pneumoniae", and search options like Web, Images, Groups, News, Froogle, Local, and Desktop.
- Cached messages:** A section below the search bar displays a cached message titled "Google Alert - antibiotic" with details like "From: Google Alerts" and "Date: Oct 26 2004 - 12:26am".
- Search term category:** A table at the bottom left lists search terms and their values, such as "Pathogen species name" with the value "Staphylococcus".

Integrated Google Desktop interface to email matches



Data Sources



Data Sources

- Local desktop
 - Documents (PDF, PowerPoint, Word, Excel, Email, etc.)
 - Data files (raw text, XML, etc.)
- Intranet
 - File servers
 - Documents (PDF, PowerPoint, Word, Excel, etc.)
 - Data files (raw text, XML, etc.)
 - Relational databases (commercial & custom)
 - Drug discovery databases
 - Reagent catalogs
 - Document management systems
- Internet
 - World-wide web
 - Specialized data sources (PubMed, USPTO)

The search paradigm allows data to reside anywhere on the network and accommodates a wide variety of data formats



Drug Discovery Database Sources

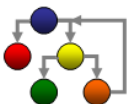
- Drug discovery data stored in relational databases (e.g., ORACLE) can be searched using SLabSearch™ through custom integration with the indexing engine
- Databases include:
 - LIMS (custom in-house and commercial)
 - Functional genomics (custom in-house and commercial, e.g., microarray databases)
 - HTS (custom in-house and commercial e.g., ActivityBase, RS3)
 - Chemistry (custom in-house and commercial, e.g., ISIS, ACCORD)
 - PK/PD (custom in-house and commercial)
 - Pre-clinical (custom in-house and commercial)

Drug discovery data resides in a wide variety of formats across the research network



Why Search Engines?

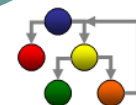
- Provides sophisticated algorithms for ranking search results relevance
- Provides the flexibility for data to reside in their native platforms (files, documents, presentations, RDBMS, web pages)
- Provides the flexibility for data to be distributed throughout the research network
- Provides application developers to customize the behavior of the search engine for handling specialized data content and user interface display
- Provides integrated security methods



Search Engines

- SLabSearch™ uses search engines to index content and provide a programmable query interface
 - Google Desktop™ – indexing local content
 - Google Mini™ or Enterprise Search Appliance™ – indexing intranet content
 - Google Web – indexing world-wide web content
 - Specialized site-specific search interfaces (e.g., PubMed, USPTO)

Google Search engines provide support complex boolean logic in forming search queries and rank search results based on document relevance. The user challenge is in specifying these boolean expressions to accurately reflect the search intent...



Conceptual Searching

- Searching for relevant data requires knowledge of the vocabulary used for naming the data and knowledge elements
- The goal of a conceptual search is to increase the *precision* and *recall* of a search by including the relevant terms and their synonyms that specify a scientific concept and formulate a *concept search expression*
- Searching by *concept* rather than individual terms provides a more precise context for the terms used in a search
- The search engine's match relevance algorithms improve the ranking of documents by incorporating more relevant terms
 - We miss things because our searches are imprecise. Relevant documents may be further down the list or integrated with irrelevant hits
 - We miss things because people use different (but accurate) terms and spellings to refer to the same thing

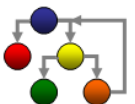
Conceptual searching is required to overcome the limitations of keyword searching



Semantic Technologies

- Ontologies are an emerging Semantic Web data representation technology that can be used for capturing the complex hierarchical relationships (semantics) between the terms used in complex terminologies
- Reasoning engines applied to ontologies can dynamically synthesize new term relationships (based on the ontology logic) that are *implied* by the rules of the terminology
- Ontologies provide a precise and portable format for concept and knowledge representation that can be exploited for constructed complex search queries

The Semantic Web is bringing to the Internet a revolutionary way of locating and aggregating information using



SLabSearch™ Search Ontologies

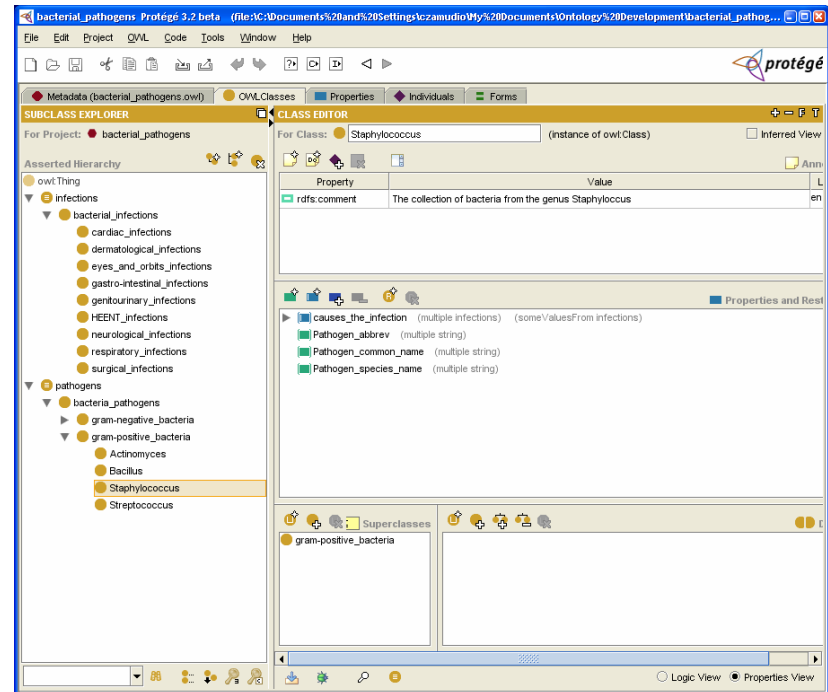
- Ontologies are used to capture the complex terminologies used in drug discovery research
- The ontologies take on the form of hierarchical logic expressions
 - From the general to specific
 - Relationships constrain the range of values that terms can take within a concept
 - Term synonyms capture the variation in the labels for concepts or concept instances

Bacterial pathogen terminology hierarchy	Pathogen term relationships to infections
<ul style="list-style-type: none">● <i>Staphylococcus aureus is a type of Staphylococci</i>● <i>Some staphylococci are a type of gram-positive pathogenic bacteria</i>● <i>Some gram-positive pathogenic bacteria are a type of pathogenic bacteria</i>● <i>Pathogenic bacteria are a type of bacteria</i>● <i>Bacteria are a type of microbial organism</i>	<ul style="list-style-type: none">● <i>Pathogenic bacteria cause infection</i>● <i>Gram-positive bacteria may cause respiratory infection</i>● <i>Staphylococcus aureus cause the respiratory infection pneumonia</i>



Ontology Development

- Graphical ontology development tools exist for capturing these complex terminologies and term relationships
- The ontology files use world-wide standard XML formats for portability



The specification of search ontologies can be performed by informatics specialists working closely with scientific research domain specialists



Public Ontology Sources

- Scientific ontologies are being organized and distributed through sites such as:
 - European Bioinformatics Institute (EBI) Ontology Lookup Service (<http://www.ebi.ac.uk/ontology-lookup/>) containing over 50 (and growing) scientific ontologies distributed using Web Services protocols
 - Open Biomedical Ontologies (<http://obo.sourceforge.net/>)

Semantic Laboratories transforms these public ontologies into a consistent “search ontology” format for incorporation into SLabSearch™



Drug Discovery Ontologies

- Ontologies provide the source of scientific terminologies used to help construct conceptual queries
- Ontology classes include, e.g.,:
 - Assay and Assay result type ontologies
 - Chemistry ontologies
 - Drug ontologies
 - Drug target ontologies
 - Disease ontologies
 - FDA submission ontologies
 - Pathogen ontologies
 - Reagent ontologies

Drug discovery ontologies will help overcome the barriers to integrating a common terminology for labeling and finding relevant information across a wide variety of disciplines



SLabSearch™ Value

- Support the re-use of research data and knowledge
- Support the re-factoring of research data
- Discover relevant data and knowledge available across the research enterprise that would otherwise be “hidden” to researchers
- Search for data by content rather than structure
- Reduce the need for constructing “centralized” data repositories (e.g., data warehouses)
- Increase the precision of data retrieval by incorporating terminologies within a scientific context through semantic technologies
- Single application to create semantic searches, save and share semantic searches and save and share search results



SLabSearch™ Configuration

- Desktop
 - Windows XP
 - Sun Microsystems Java Runtime Environment 1.5 (free from Sun Microsystems)
 - Microsoft Internet Explorer
 - Microsoft Office
 - Google Desktop™ (free from Google)
- Server (e.g., Linux)
 - ORACLE XE (free from Oracle)
 - Google Mini™ (up to 300,000 documents) or Google Enterprise Search Appliance™ (up to 15 million documents)

