# Slicing through the Scientific Literature

Christopher Baker (1), Patrick Lambrix (2), Jonas Laurila Bergman (2), Rajaraman Kanagasabai (3), Wee Tiong Ang (3)

(1) University of New Brunswick, (2) Linköpings Universitet,
(3) ASTAR Singapore

# Literature search

- Huge amount of scientific literature.
- Need to integrate a spectrum of information to perform a task.

# Literature search

- How to know what is in the repository
  - Lack of knowledge of the domain
- How to compose an expressive query
  - Lack of knowledge of search technology

# Example scenario

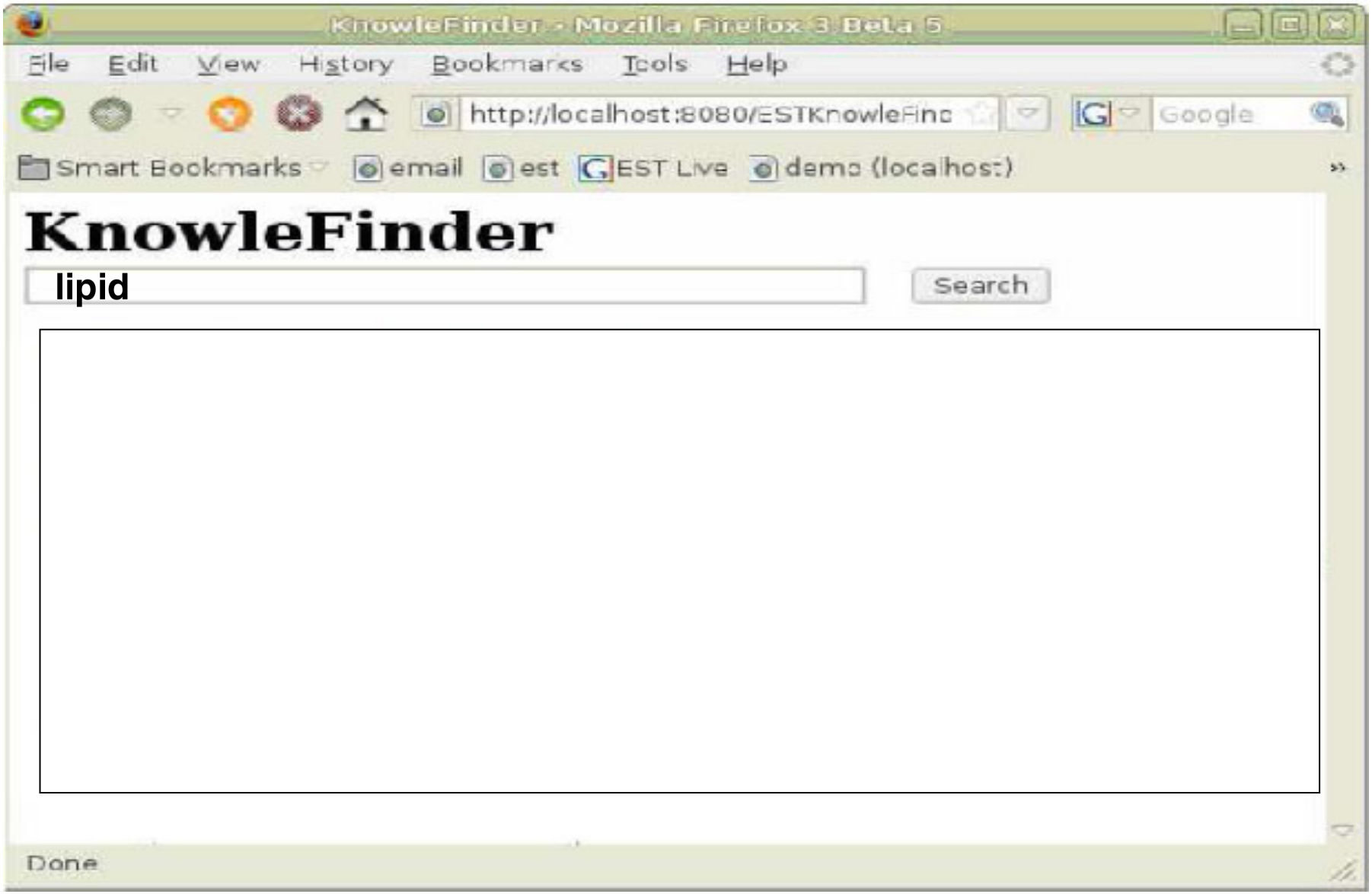
“Lipid”

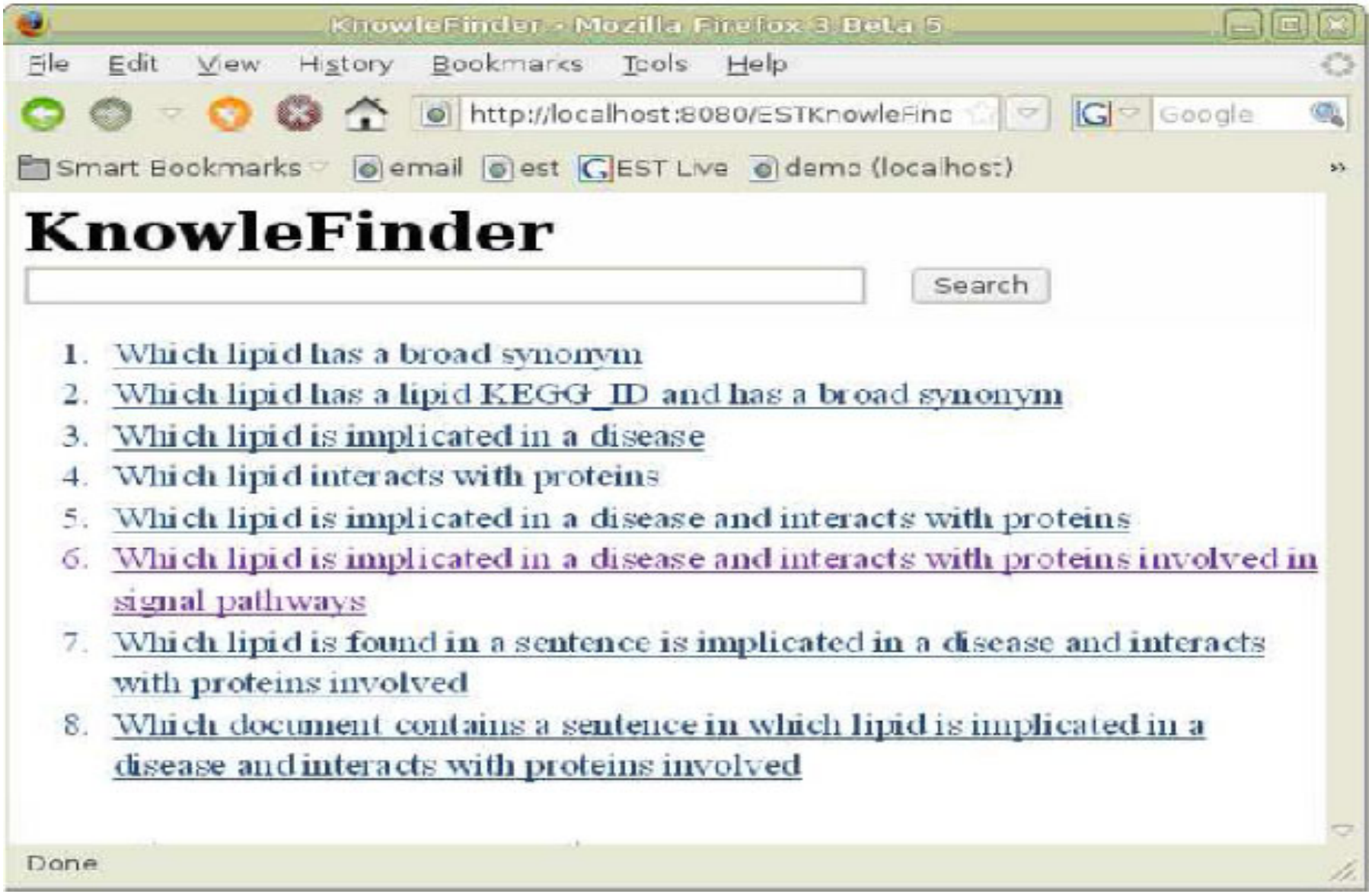
- Keyword search returns all documents containing lipid.
  - No knowledge; terminology problem
- Relationships: use of multiple keywords with/without boolean operators, e.g. *lipid and disease*

# Example scenario

“Lipid”

- Keyword search returns a list of relevant questions concerning lipid. User selects question and retrieves knowledge and provenance documents.
- Multiple search terms: requirement that there are relevant connections between the keywords.


KnowleFinder - Mozilla Firefox 3 Beta 5
File Edit View History Bookmarks Tools Help
http://localhost:8080/ESTKnowleFinc
Google
Smart Bookmarks
email
est
EST Live
demo (localhost)
KnowleFinder
lipid
Search
Done


KnowleFinder - Mozilla Firefox 3 Beta 5
File Edit View History Bookmarks Tools Help
http://localhost:8080/ESTKnowleFinc
Google
Smart Bookmarks
email
est
EST Live
demo (localhost:)
KnowleFinder
Search
1. Which lipid has a broad synonym
2. Which lipid has a lipid KEGG_ID and has a broad synonym
3. Which lipid is implicated in a disease
4. Which lipid interacts with proteins
5. Which lipid is implicated in a disease and interacts with proteins
6. Which lipid is implicated in a disease and interacts with proteins involved in signal pathways
7. Which lipid is found in a sentence is implicated in a disease and interacts with proteins involved
8. Which document contains a sentence in which lipid is implicated in a disease and interacts with proteins involved
Done

## Question

NLG: Which lipid is implicated in a disease and interacts with proteins involved in signal pathways ?

## Result

| Protein | Lipid | Disease | Signal Pathway |
|---|---|---|---|
| | | | |
| P53 | Unsat. Fatty Acid | Ovarian Cancer | Apoptosis |

# Example scenario

- Requirements
  - Natural language interface
  - Ontology-driven query
  - Context of query terms
  - Cross-domain queries

# Outline

- Relevant queries
- Framework for slicing through the scientific literature
- Algorithms and example
- Conclusion & Future Work

# Outline

- **Relevant queries**
- Framework for slicing through the scientific literature
- Algorithms and example
- Conclusion & Future Work

# Ontologies

"*Ontologies define the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary.*"

# Ontologies

**GENE ONTOLOGY (GO)**

immune response
- i- acute-phase response
- i- anaphylaxis
- i- antigen presentation
- i- antigen processing
- i- cellular defense response
- i- cytokine metabolism
  - i- cytokine biosynthesis
    - synonym cytokine production
    - …
  - p- regulation of cytokine biosynthesis
  - …
- …
- i- B-cell activation
  - i- B-cell differentiation
  - i- B-cell proliferation
- i- cellular defense response
- …
- i- T-cell activation
  - i- activation of natural killer cell activity
  - …

We will assume a graph representation.

# Relevant queries

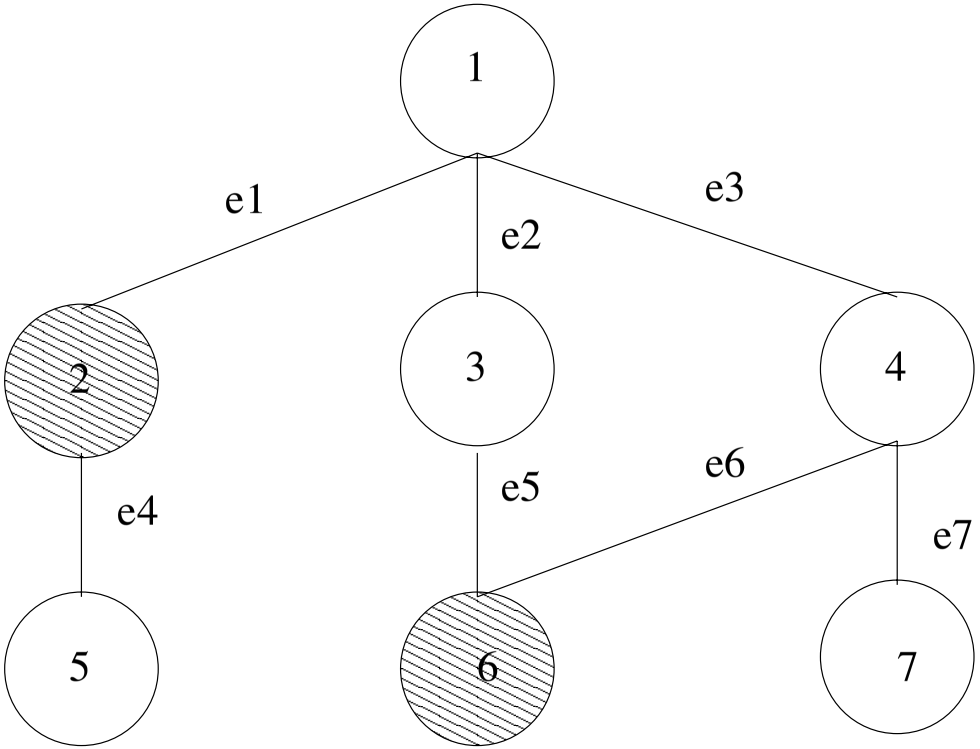
- Relevant query including a number of concepts and relations from an ontology

  connected sub-graph of the ontology that includes the concepts and relations.

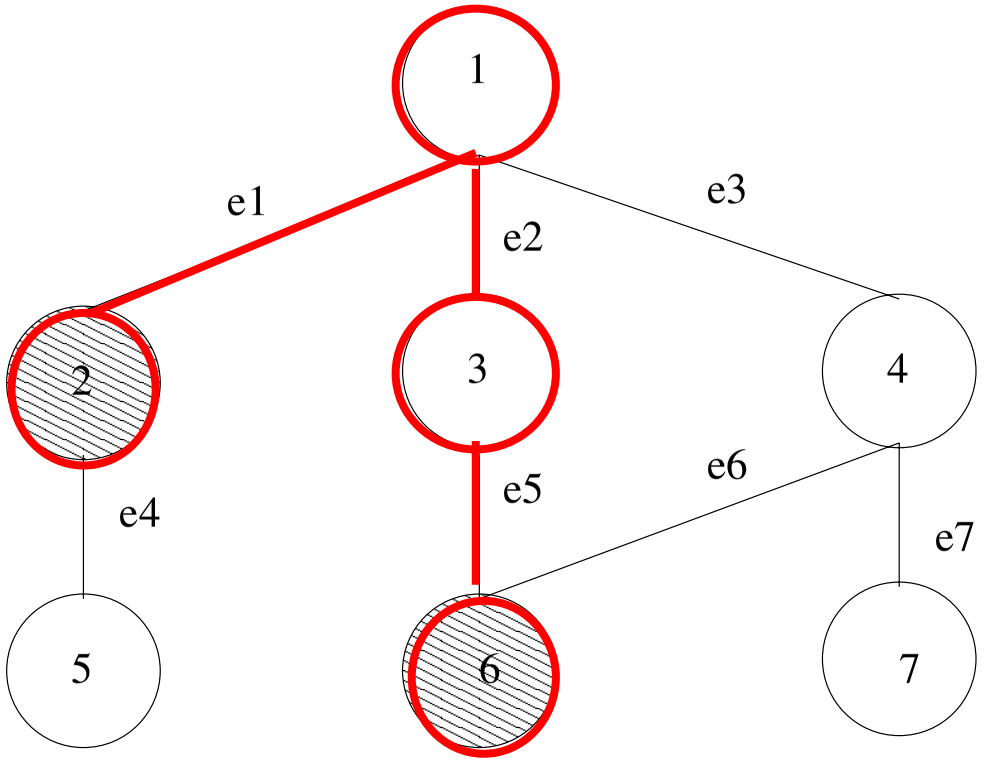
  *(query graph based on the concepts and relations;*

  *slice is set of all query graphs based on the concepts and relations)*

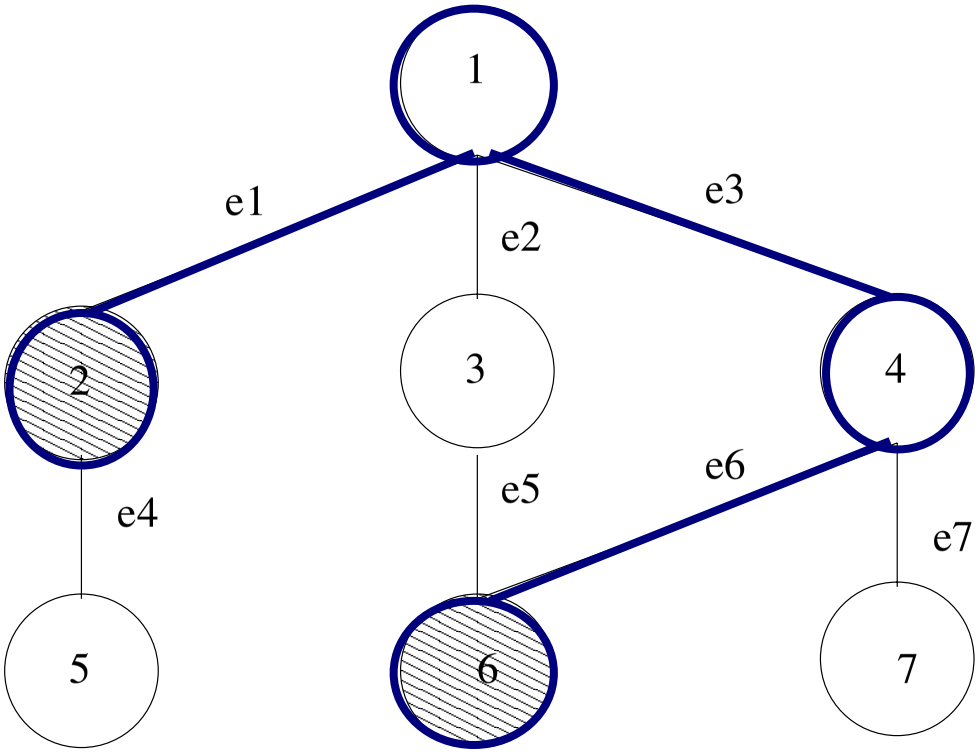
# Query graph



# Query graph



# Query graph



# Special cases

- No relations, several concepts
  - Relevant queries regarding concepts; relations are suggested by the system.
  - Difference with traditional techniques: extra requirement that search terms need to be connected in the ontology.
- No relations, one concept
  - Relevant queries including a specific query term.
  - Computes the ontological environment of the query term.

# Multiple ontologies

**GENE ONTOLOGY (GO)**

immune response
- i- acute-phase response
- i- anaphylaxis
- i- antigen presentation
- i- antigen processing
- i- cellular defense response
- i- cytokine metabolism
  - i- cytokine biosynthesis
    synonym cytokine production
    …
  - p- regulation of cytokine biosynthesis
  …

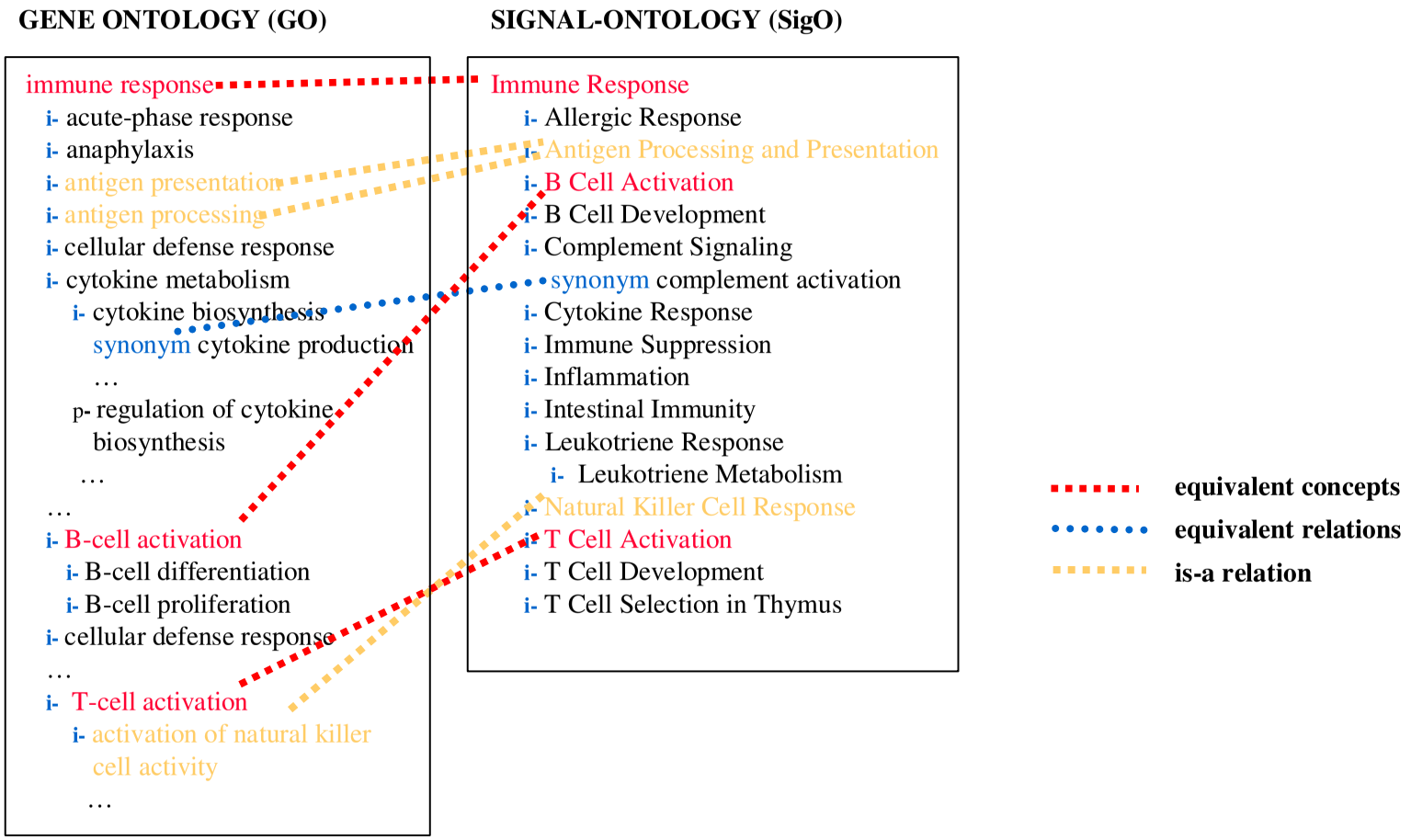
…
- i- B-cell activation
  - i- B-cell differentiation
  - i- B-cell proliferation
- i- cellular defense response

…
- i- T-cell activation
  - i- activation of natural killer cell activity
  …

**SIGNAL-ONTOLOGY (SigO)**

Immune Response
- i- Allergic Response
- i- Antigen Processing and Presentation
- i- B Cell Activation
- i- B Cell Development
- i- Complement Signaling
  synonym complement activation
- i- Cytokine Response
- i- Immune Suppression
- i- Inflammation
- i- Intestinal Immunity
- i- Leukotriene Response
  - i- Leukotriene Metabolism
- i- Natural Killer Cell Response
- i- T Cell Activation
- i- T Cell Development
- i- T Cell Selection in Thymus

# Ontology Alignment




Alignment is a set of mappings between terms in the ontologies.

# Relevant queries – multiple ontologies

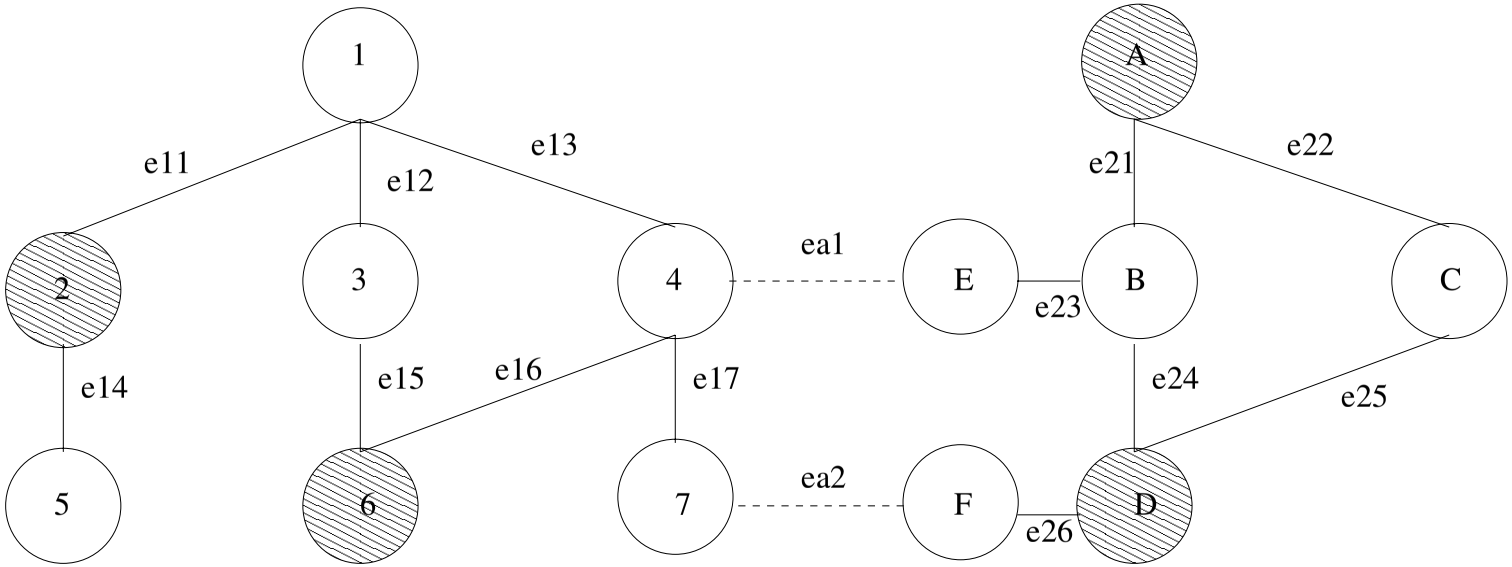
- Relevant query including a number of concepts and relations from multiple ontologies

  Query graphs connected by a path going through a mapping in the alignment.

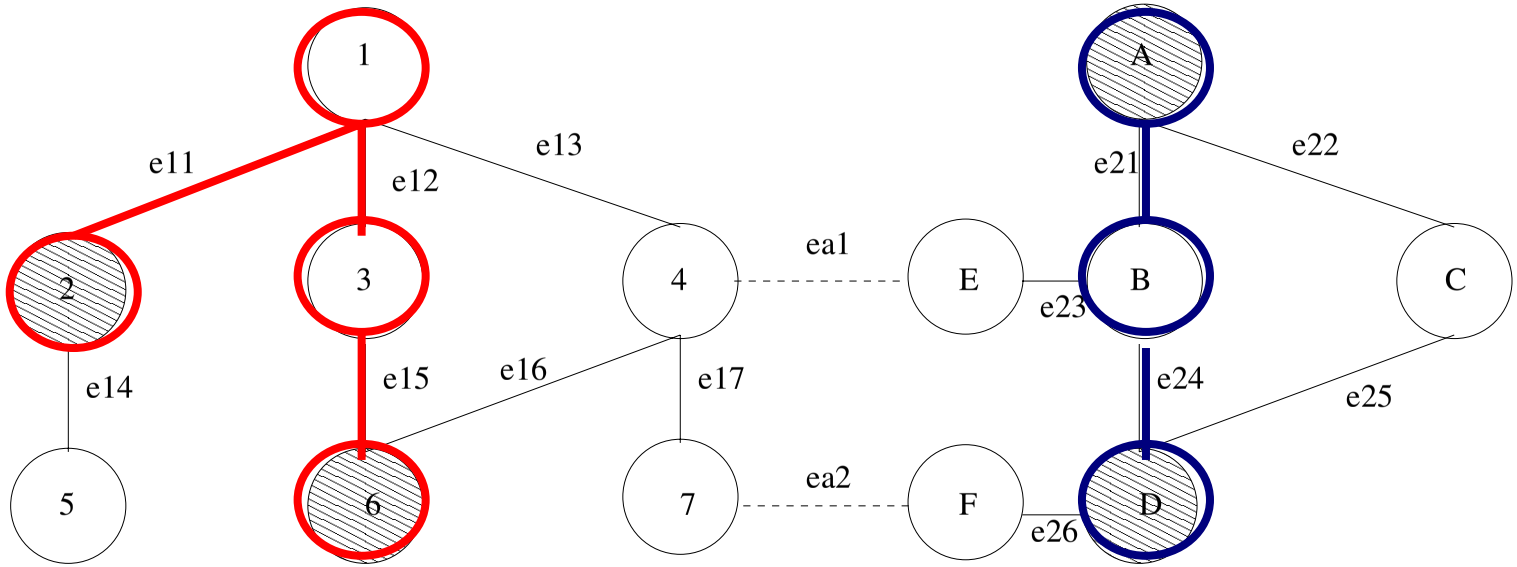
  *(aligned query graph based on query graphs;*

  *aligned slice is set of all aligned query graphs based on the query graphs)*

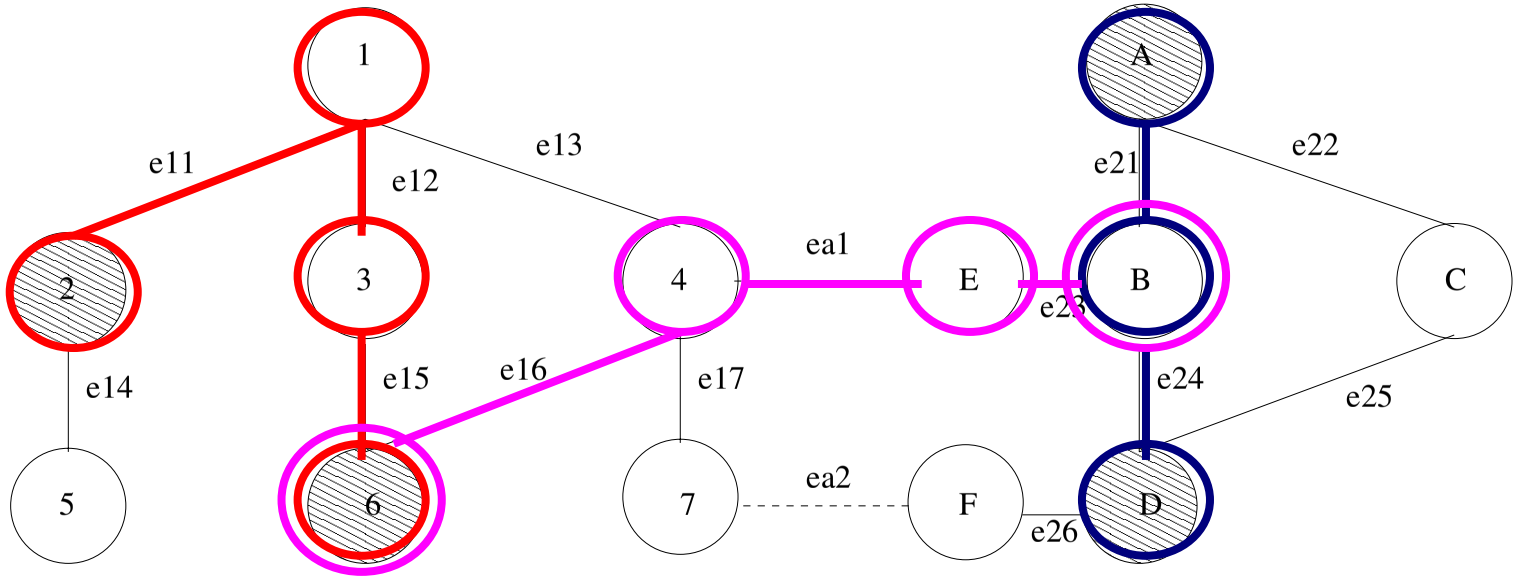
# Aligned query graph



# Aligned query graph



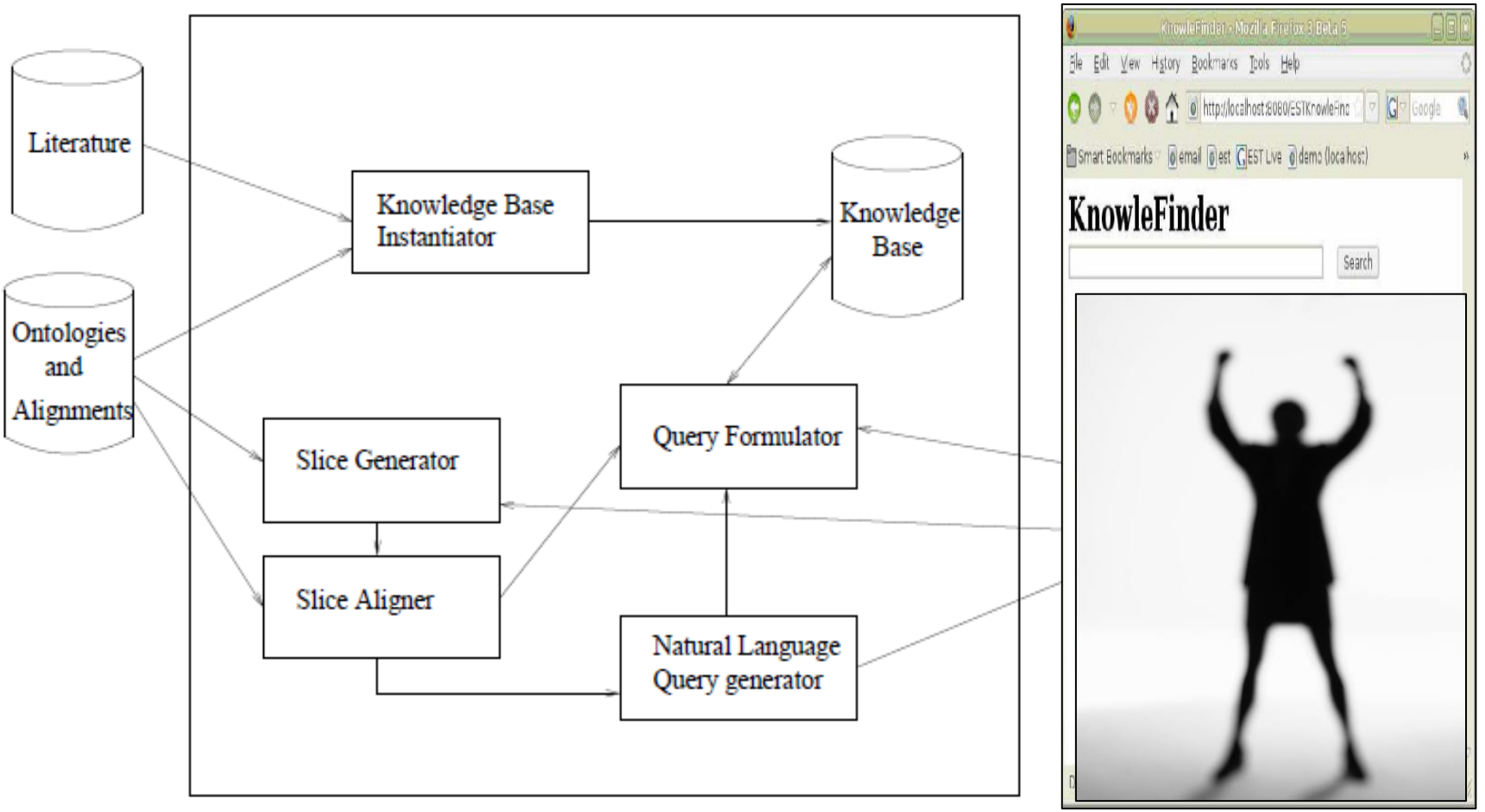
# Aligned query graph



# Outline

- Relevant queries
- **Framework for slicing through the scientific literature**
- Algorithms and example
- Conclusion & Future Work

# Framework


Literature
Ontologies and Alignments
Knowledge Base Instantiator
Knowledge Base
Slice Generator
Slice Aligner
Query Formulator
Natural Language Query generator
KnowleFinder
Search

# Framework

- External resources
  - Literature document base
  - Ontology and ontology alignment repository
- Computed resources
  - Knowledge base

# Framework

- Process and translation
  - Knowledge base instantiation
  - Slice generation and alignment
  - Natural language query generation

- Query

# Outline

- Relevant queries
- Framework for slicing through the scientific literature
- **Algorithms and example**
- Conclusion & Future Work

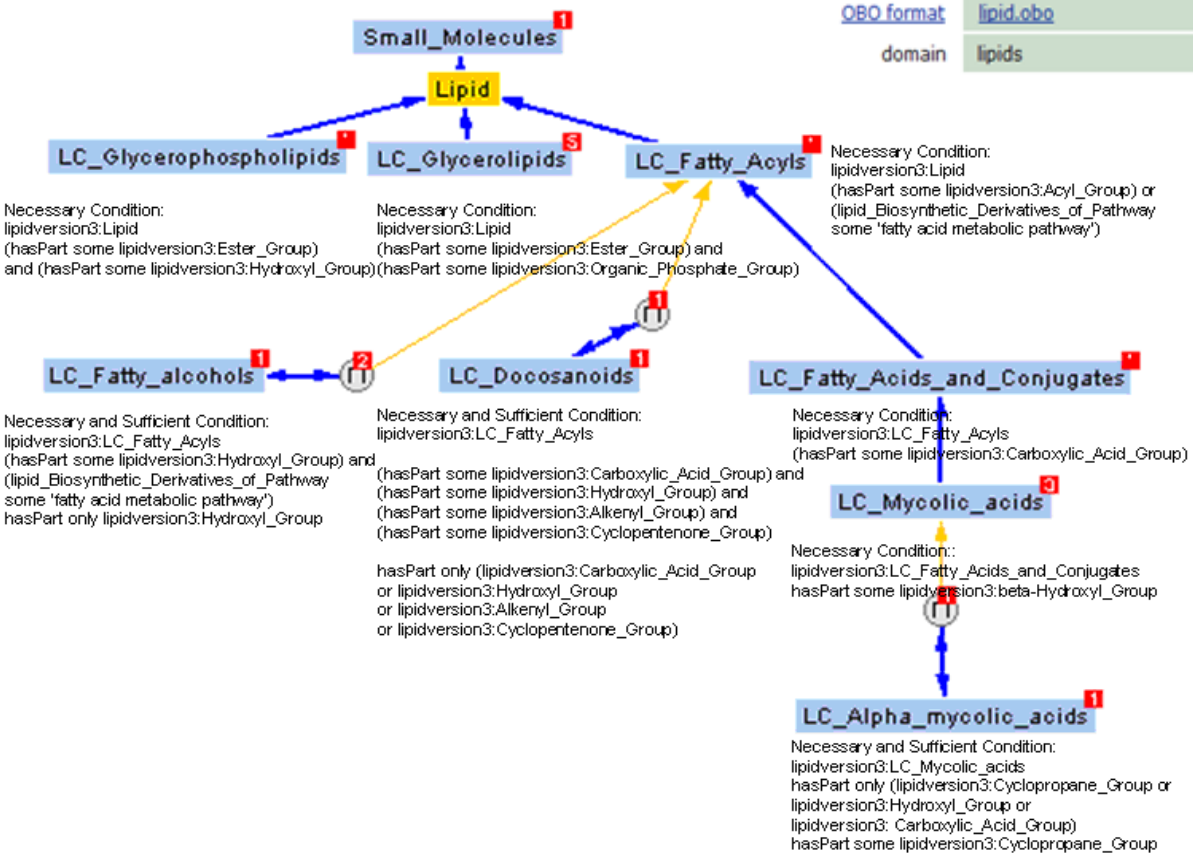
# External resources

- Literature document base
  - Generated from a collection of 7498 PubMed abstracts relevant for Ovarian Cancer. 683 papers included lipid names from which 241 full papers were downloadable.
- Ontology and ontology alignment repository
  - Lipid ontology
  - Signal ontology
  - Aligment using SAMBO

## *Lipid Ontology*

Lipid research is increasingly integrated within systems level biology such as lipidomics where lipid classification is required before appropriate annotation of chemical functions can be applied. The ontology describes the LIPIDMAPS nomenclature classification explicitly using description logics (OWL-DL). Lipid classes are organized hierarchically with the super-classes restricted by generic necessary conditions. More specific necessary conditions are used to define membership requirements for sub classes of lipid according to appropriate functional groups.

| | |
|---|---|
| namespace | LiPrO |
| current activity | Active |
| contact | Christopher Baker |
| OWL format | LiPrO.owl |
| OBO format | lipid.obo |
| domain | lipids |

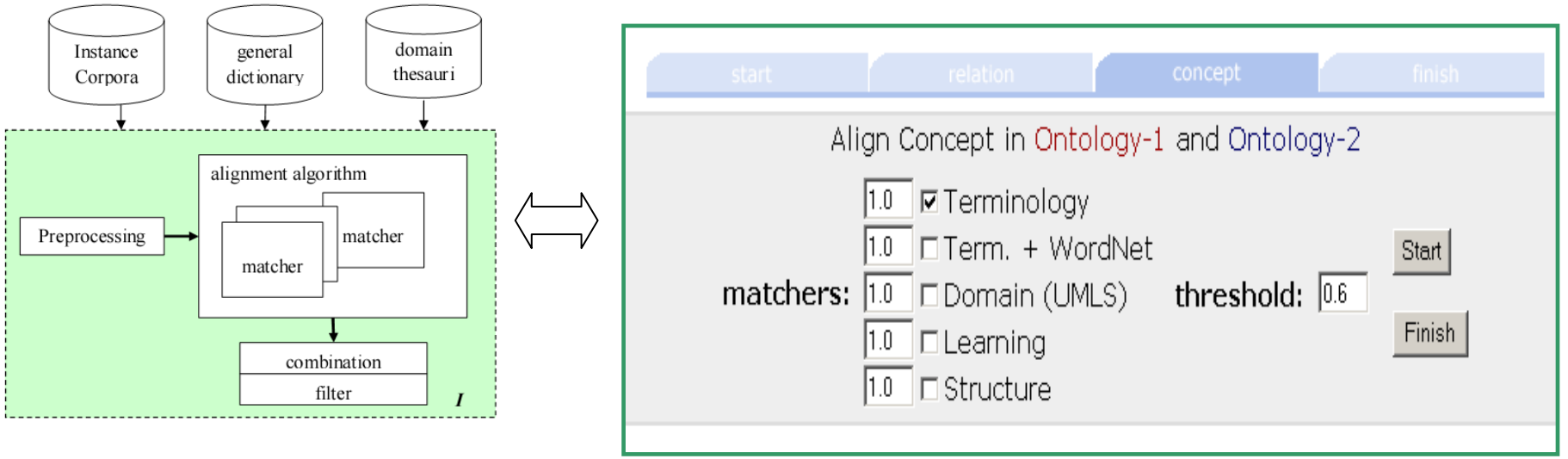


# SAMBO (1)

- SAMBO (System for Aligning and Merging Biomedical Ontologies)
  - Phase I

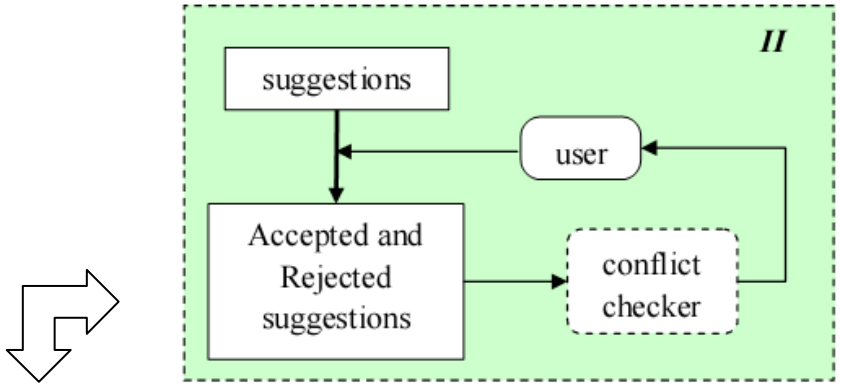
*Winner Anatomy Track of OAEI 2008*

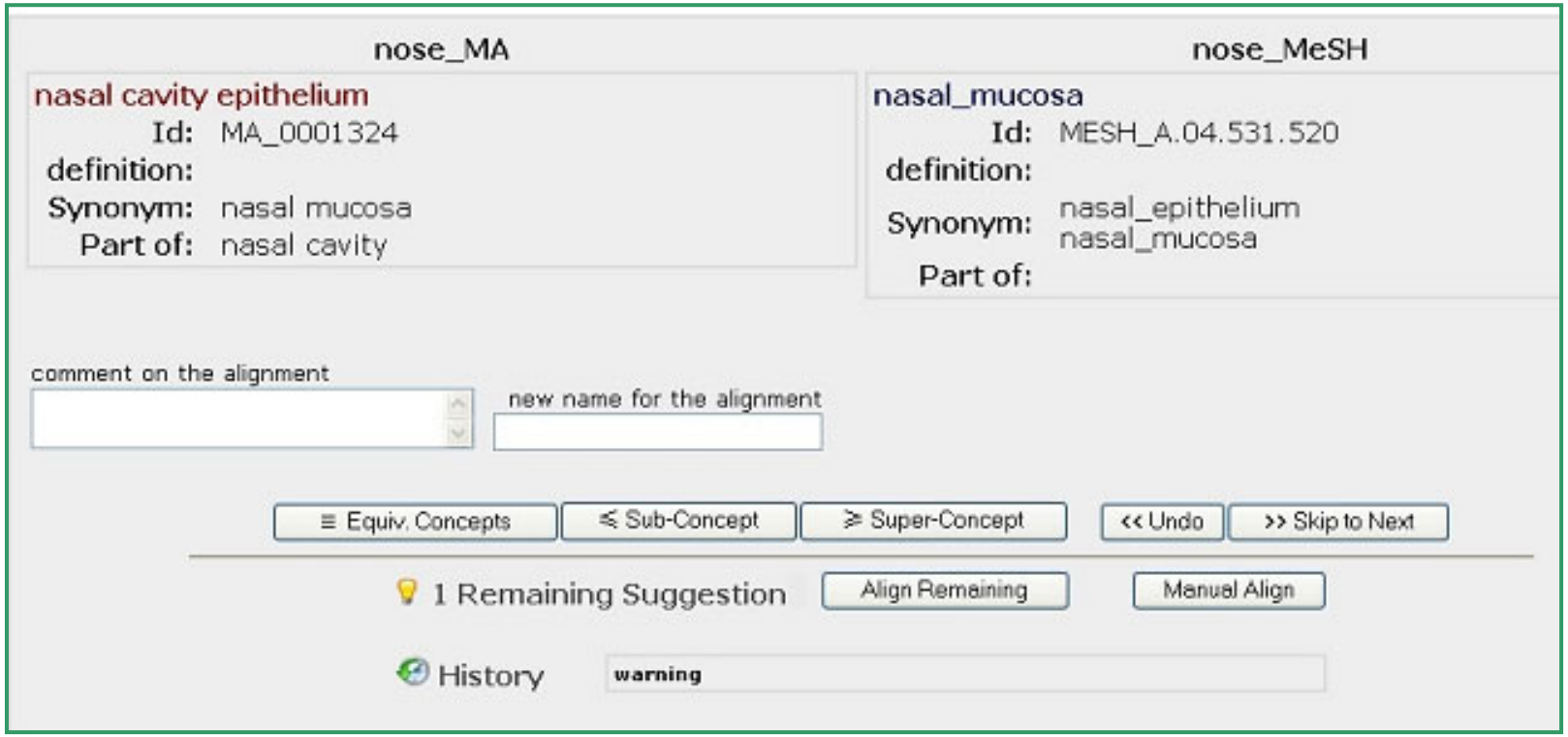
    - Matchers
    - Weighted sum combination of matcher results
    - Single threshold filtering



# SAMBO (2)

- Phase II:



# Knowledge base instantiation

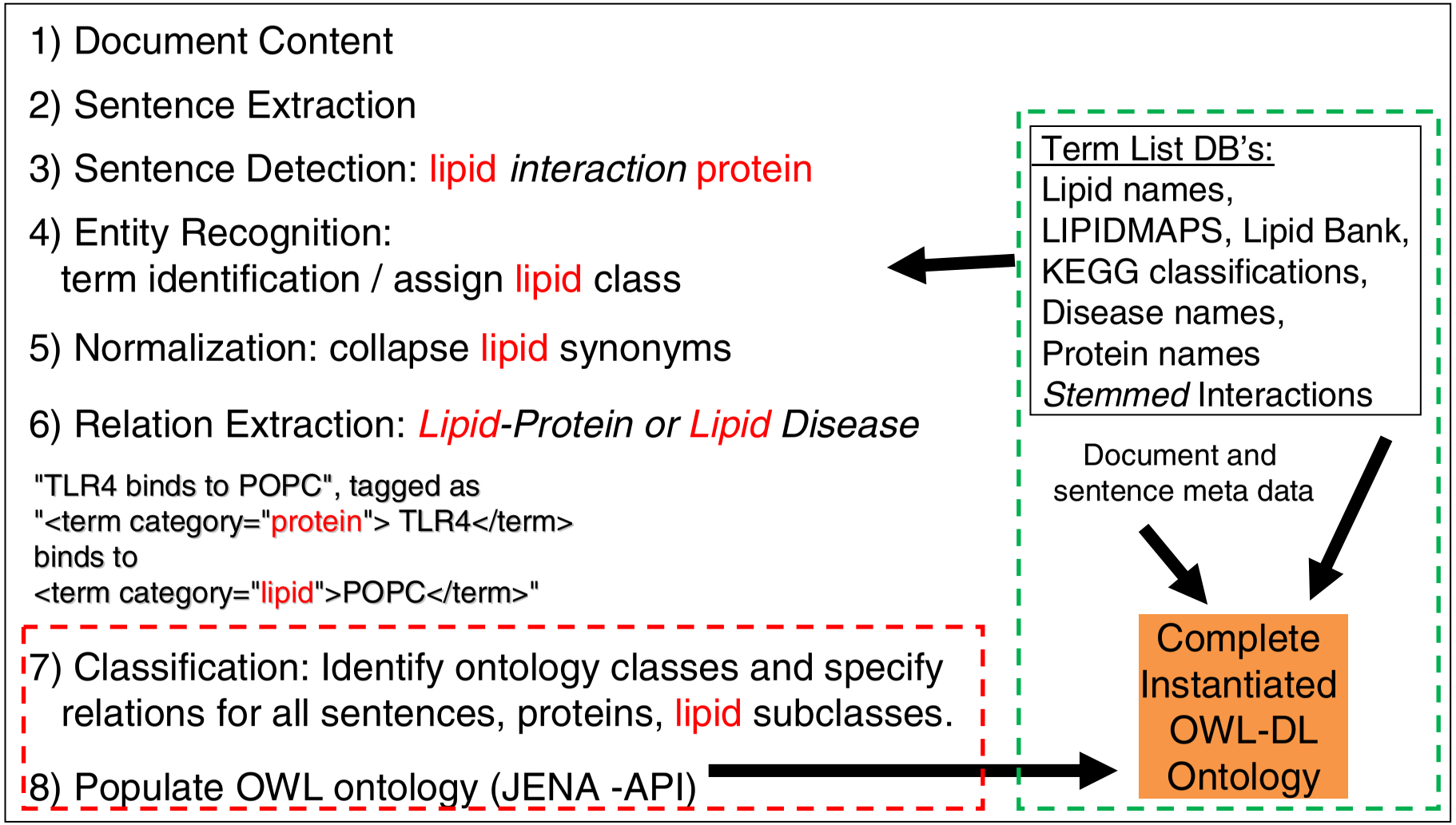
BioText toolkit

- Named entity recognition
  - Gazetteer approach
  - Termlists
    - Lipids: LIPIDMAPS, LipidBank, KEGG, IUPAC
    - Proteins: Swiss-Prot
    - Diseases: disease ontology of Center of Genetic Medicine
  - Tags found entities with ontology concepts

- Normalization and grounding

- Relation detection
  - Based on co-occurrence in sentence
    + rule set developed with domain expert

# Knowledge base instantiation



# Mined Interactions

**Table 1.** Interactions mined from the ovarian cancer bibliome. OC and AP represent a cancer and apoptosis pathway proteins respectively.

| **Interaction Type** | **Abstract** (7498) | **Full Paper** (241) |
|---|---|---|
| OC-OC | 223 | 13 |
| OC-AP | 505 | 195 |
| OC-Lipid | 11 | 14 |
| OC-Hormone | 8 | 1 |
| AP-AP | 113 | 59 |
| AP-Lipid | 10 | 8 |
| Lipid-Lipid | 3 | 23 |
| Lipid Hormone | 2 | 18 |
| Protein Hormone | 9 | 2 |
| Hormone-Hormone | 2 | 6 |

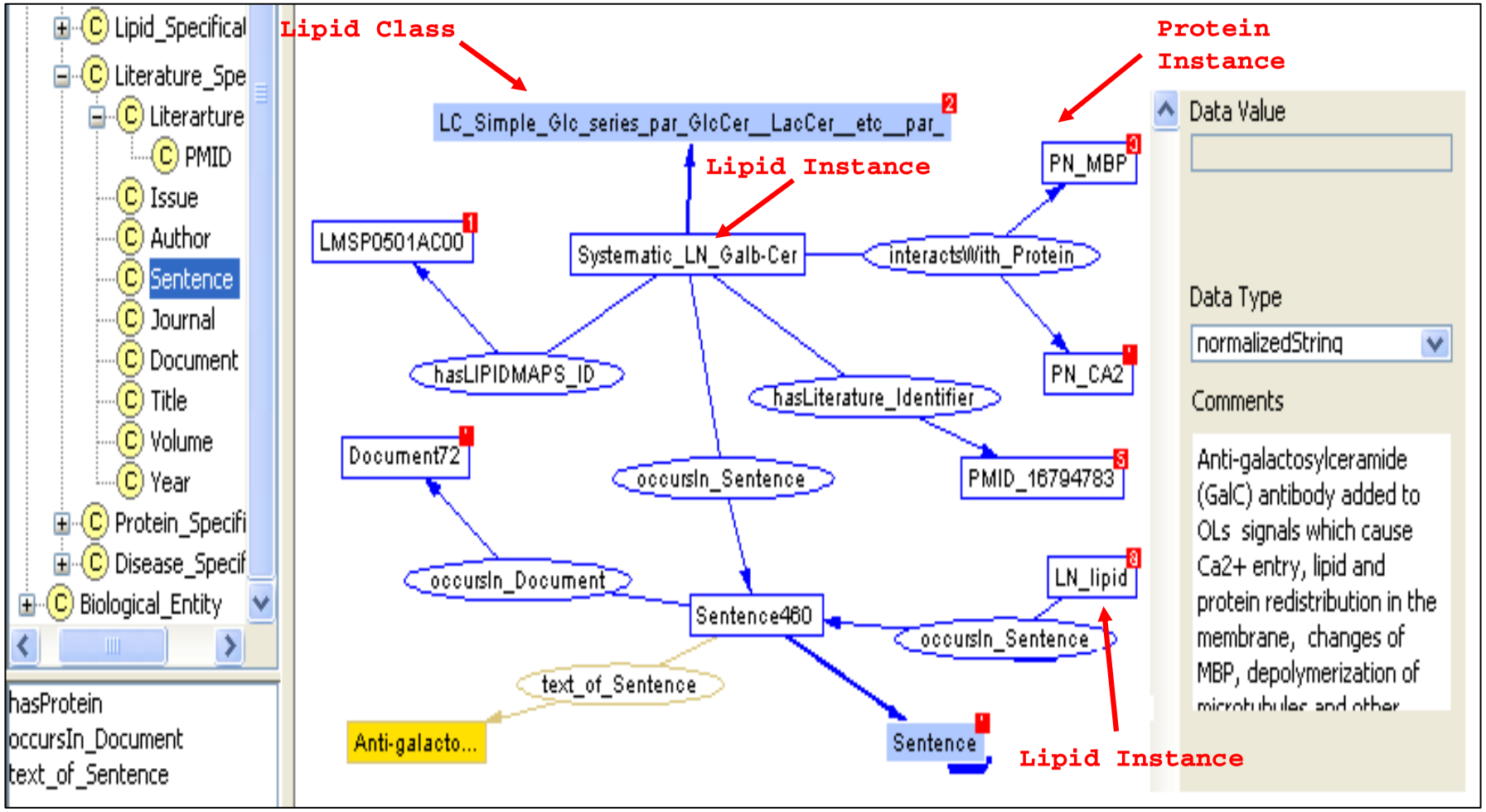
# Knowledge base instantiation

- Population
  - Concepts
  - Properties

Knowledge

Provenance documents

# Knowledge base instantiation



# Slice generation

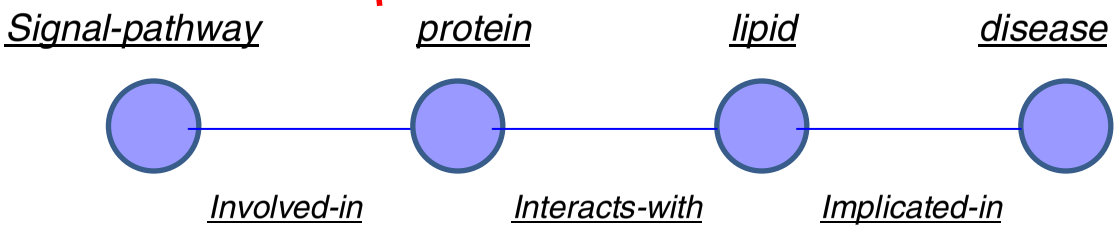
- Current implementation focuses on slices based on concepts.
- Depth-first traversal of ontology to find paths between given concepts; paths can be put together to find slices/query graphs.

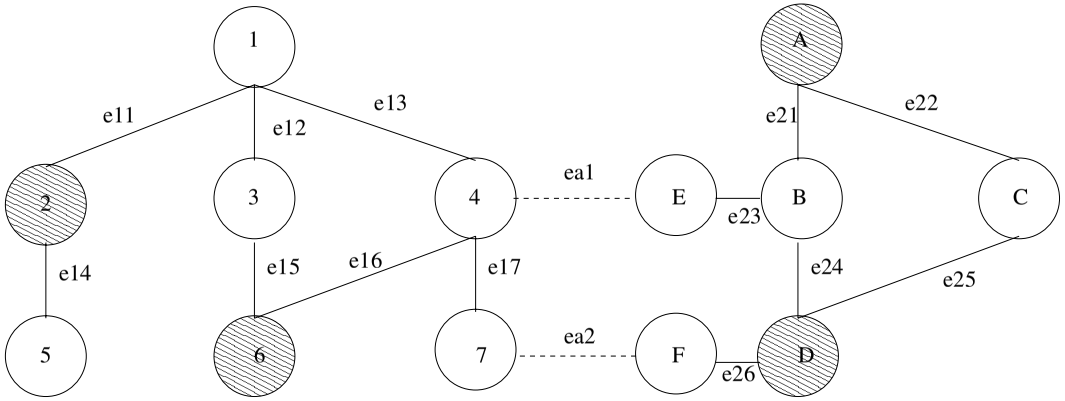
# Slice alignment

- Algorithm computes subset of aligned slice.
- Assumption: shorter paths represent closer relationships.
- Algorithm connects slices using shortest paths from given concepts in one ontology to given concepts in other ontology.

# Slicing through the literature

```
nRQL: (RETRIEVE (?X ?Y ?Z ?W)
(AND (?X Protein) (?Y Lipid) (?Z Disease) (?W SignalPathway)
(?X ?Y Interacts_with)(?Y ?Z Implicated_in)(?X ?W Involved_in)))
```



# Natural language query generation

- Triple representation:

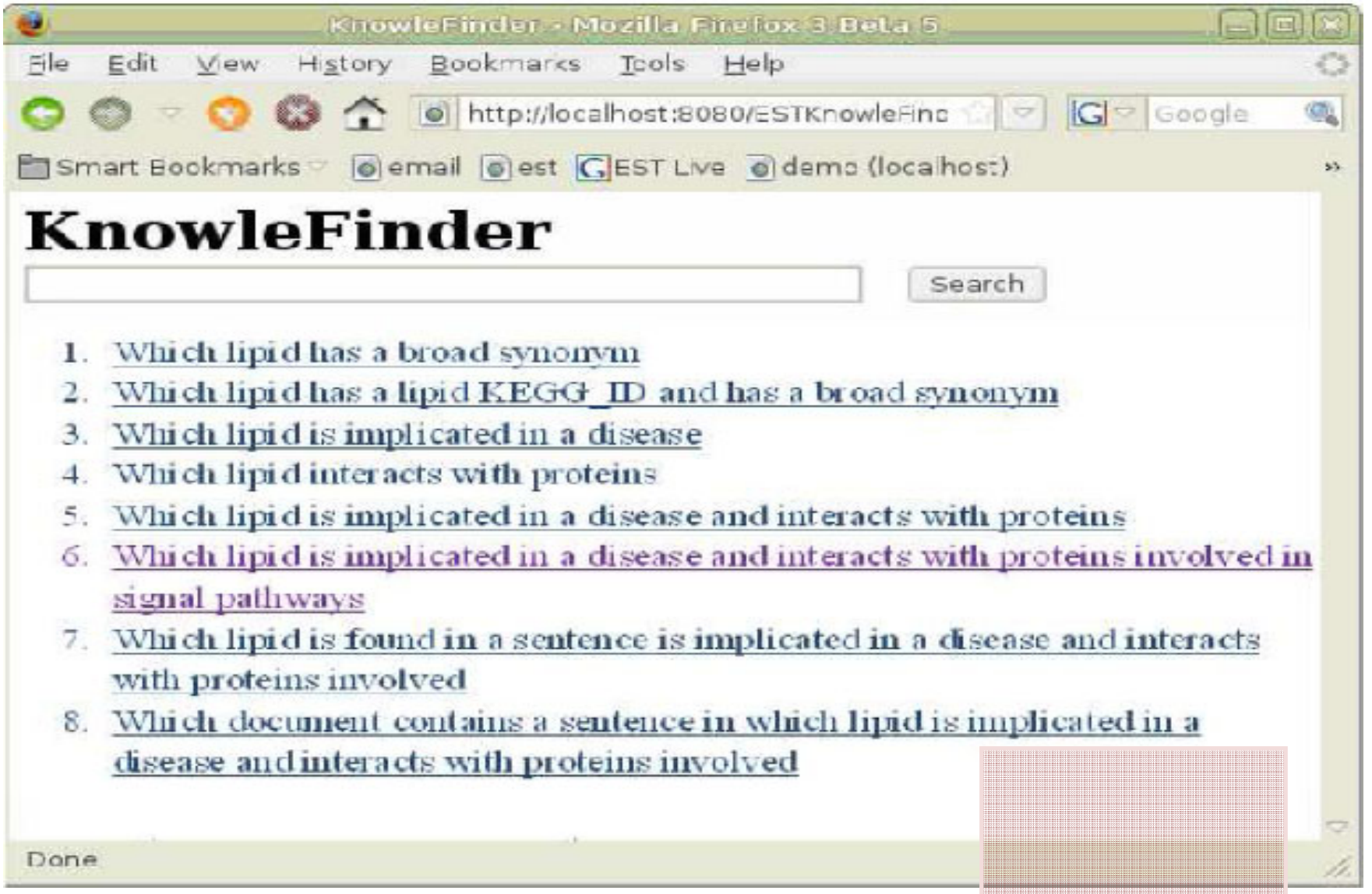
  *<lipid, interacts-with, protein>*

- Rule base to generate NL statements.

  *What lipid interacts with proteins?*

  - Learned from examples.

- Aggregation of statements from different triples, grammar checking.


KnowleFinder - Mozilla Firefox 3 Beta 5
File Edit View History Bookmarks Tools Help
http://localhost:8080/ESTKnowleFinc
Google
Smart Bookmarks
email
est
EST Live
demo (localhost:)
KnowleFinder
Search
1. Which lipid has a broad synonym
2. Which lipid has a lipid KEGG_ID and has a broad synonym
3. Which lipid is implicated in a disease
4. Which lipid interacts with proteins
5. Which lipid is implicated in a disease and interacts with proteins
6. Which lipid is implicated in a disease and interacts with proteins involved in signal pathways
7. Which lipid is found in a sentence is implicated in a disease and interacts with proteins involved
8. Which document contains a sentence in which lipid is implicated in a disease and interacts with proteins involved
Done

# Query

- Send nRQL query to RACER.

**Question**

NLG: Which lipid is implicated in a disease and interacts with proteins involved in signal pathways ?

```
nRQL:(RETRIEVE (?X ?Y ?Z ?W)
 (AND (?X Protein) (?Y Lipid) (?Z Disease) (?W SignalPathway)
 (?X ?Y Interacts_with)(?Y ?Z Implicated_in)(?X ?W Involved_in)))
```

**Result**

| Protein | Lipid | Disease | Signal Pathway |
|---|---|---|---|
| | | | |
| P53 | Unsat. Fatty Acid | Ovarian Cancer | Apoptosis |

# Outline

- Relevant queries
- Framework for slicing through the scientific literature
- Algorithms and example
- **Conclusion & Future Work**

# Conclusions

- Framework for literature search dealing with
  - Lack of knowledge of the domain.
  - Lack of knowledge of the search technology.

- Proof of concept implementation.

# Future Work

- Heuristics and their influences for slice generation.
- Tradeoff in query generation between completeness and information overload.
- Relevance measure and query ranking.
- Optimization of slice generation.

# Future Work

- Integration in larger system
- Integrated implementation.
- Scalability testing.