Literature-driven, Ontology-centric Knowledge Navigation for Lipidomics

#Rajaraman Kanagasabai[1], #Hong-Sang Low[2], Wee Tiong Ang[1], Anitha Veeramani[1], Markus R. Wenk[2], Christopher J. O. Baker*[1],

[1] Data Mining Department, Institute for Infocomm Research, Singapore.
cbaker@i2r.a-star.ediu.sg
[2] Department of Biochemistry and Department of Biological Sciences, Centre for Life Sciences, Singapore.

As the semantic web vision continues to proliferate a gap still remains in the full scale adoption of such technologies. The exact reasons for this continue to be the subject of ongoing debate, however, it is likely the emergence of reproducible infrastructure and deployments will expedite its adoption. We illustrate the recognizable added value to life science researchers gained through the convergence of existing and customized semantic web technologies (content acquisition pipelines supplying legacy unstructured texts, natural language processing, OWL-DL ontology development and instantiation, reasoning over A-boxes using a visual query tool). The resulting platform allows lipidomic researchers to rapidly navigate large volumes of full-text scientific documents according to recognizable lipid nomenclature, hierarchies and classifications. Specifically we have enabled searches for sentences describing lipid-protein and lipid-disease interactions.

1 Introduction

A series of existing technologies are now recruited along with semantic technologies to build scientific information systems delivering enriched value-added performance [1]. In particular there is an increasing need to link relevant content to semantic web infrastructure either by tagging existing web content and linking it to semantic metadata [2] or by indexing / summarizing legacy formats using algorithms focused on raw text analysis. In this latter case, where NLP approaches are now well established there would appear to be a complementary fit. Specifically the results of text analysis such as marked up text segments, which are typically deposited in relational databases, can be repurposed as instances to precisely defined concepts in ontologies. Likewise the relations between such named entities in text segments can also be instantiated to knowledge-bases. Such knowledgebases can represent a searchable summary of large volumes of literature [3]. Ontologies can provide richly cognitive query models to instantiated knowledgebases and in conjunction with reasoning engines can facilitate instance retrieval for knowledge discovery tasks. Here we focus on a contemporary application domain, Lipidomics, with the goal of

building an ontology-centric navigation platform to facilitate knowledge discovery for life scientists.

In section 2 we describe the architecture supporting the platform. In section 3 we introduce the *status quo* and current challenges in lipid research motivating for the development of the lipid ontology, which we also describe. In section 4 we describe the content acquisition strategy, natural language processing and the lipid-specific ontology instantiation strategy. In section 5 we describe the features of the knowledge navigator interface, discuss user scenario and query paradigms for interrogating the scientific literature.

2 Ontology-centric Content Delivery Platform

The outline of our platform is shown in Figure 1. It comprises of a content acquisition engine that drives the delivery of literature. This engine takes user keywords and retrieves full-text research papers from distributed public repositories and converts them to a custom format ready for text mining. A workflow of natural-language processing algorithms identifies target concepts or keywords and tags individual sentences according to the terms they contain. Sentences are instantiated (as A-boxes) using a custom designed java program to the ontology's literature specification (sentence concept) and relations to instances of each target concept found in the sentence are added into the ontology. The fully instantiated ontology is reasoned over using the reasoning engine RACER and it's A-box query language nRQL [4]. A custom built visual query interface, described in section 5, facilitates query navigation over instantiated object properties and visualization of datatype properties in the ontology.

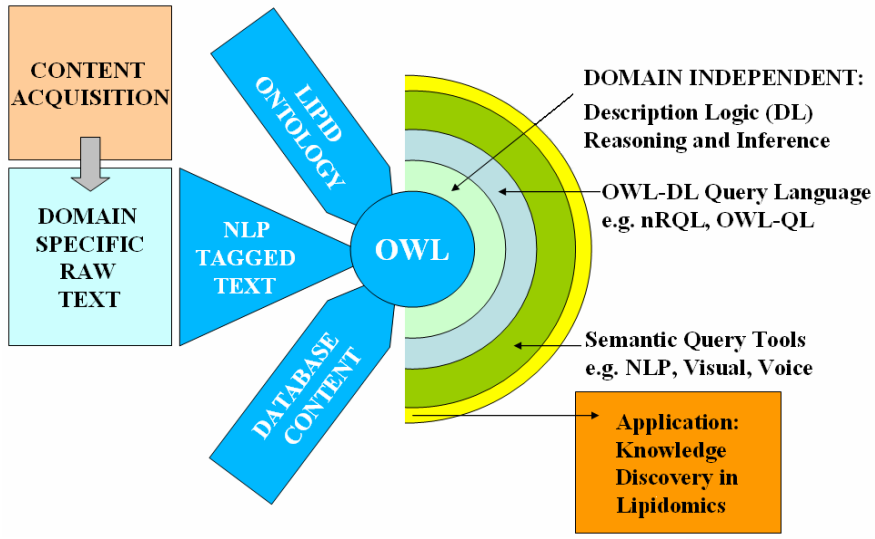

Fig. 1. Ontology-centric knowledge navigation system architecture

3

3 Lipids and Lipidomics

Lipids and their metabolites have a very crucial role in the biology and cellular functions of many living organisms. They are used for energy storage, serve as the structural components of cell membranes, and constitute important signaling molecules. Consequently lipids play diverse and important roles in nutrition and health: Imbalance or abnormality in lipid metabolism often accompanies diseases such as Alzhemer's syndrome, hypercholesterolemia and cancer. Lipidomics [5] is an emerging biomedical research field with important applications in the development of drugs and biomarkers for diseases e.g. cancer and diabetes. In order to attain a better understanding of the role of lipids in physiological processes, scientists use high throughput technology in the analysis of lipid composition of living organisms. Lipidomics generates large amounts of chemical, biological, analytical data that need to be integrated and analyzed in a systematic manner. A major challenge in this regard is the lack of consistent classification for lipids.

3.1 Lipid Classification Challenges

Lipids, unlike their protein counterparts, do not have a systematic classification and nomenclature that is widely adopted by biomedical research community. To address this problem, IUPAC-IUBMB [6] developed a standardized, systematic nomenclature for lipids. The IUPAC nomenclature suffers, however, from several drawbacks. Firstly, it has not gained widespread adoption since the systematic naming of lipids according to their structures can become long and cumbersome. Furthermore the IUPAC naming scheme was often misunderstood by scientists leading to the generation of many pseudo-IUPAC names that are neither chemically or scientifically sound. Given that the IUPAC naming scheme emerged in 1976, the naming scheme has not evolved since then to accommodate the large number of novel lipid classes that have been discovered in the last 3 decades.

In this context different lipid research groups developed their own classifications of lipids which are usually very narrow and only sound for a restricted lipid category. As a result, the same lipid molecule can be classified in many different ways, and be placed under different types of classification hierarchy. A single lipid can be associated with a plethora of synonyms. Furthermore, most of these classification systems are not scientifically sound and hence, create a lot of problems for the systematic analysis of lipids.

The LIPIDMAPS consortium [7] recently developed a scientifically robust and comprehensive chemical representation and classification system that incorporates a consistent nomenclature that is closely aligned to IUPAC nomenclature yet extensible to include new lipids without a systematically defined IUPAC name. Adoption of this standard has been gradual and many research groups still use synonyms or old names. More importantly legacy literature resources predominantly contain instances of lipid synonyms not yet linked to the LIPIDMAPS systematic name or any chemically sound classification.

3.2 Lipid Ontology

It is with the above mentioned problems in mind, we developed the Lipid Ontology. The rationale behind the Lipid Ontology is manifold: (i) it serves to connect the pre-existing/legacy lipid synonyms found in literature or other databases to the LIPIDMAPS classification system; (ii) it serves as a data model to manage information on lipid molecules, define features and declare appropriate relations to other biochemical entities i.e. proteins, diseases, enzymes and pathways; (iii) it serves as an integration and query model for one or more data warehouses of lipids information (iv) it serves as a flexible and accessible format for defining the current systematic classification of lipids and lipid nomenclature, which is particularly relevant to the discovery of new lipids and lipid classes that have yet to be systematically named. The ontology currently has a total of 668 concepts and 74 properties.

The Lipid ontology emerged from a data-warehouse schema developed [8] to house lipid information and lipidomics data. Consequently the ontology inherited certain features of the data model. Information about individual lipid molecules is modeled under the Lipid and Lipid Specification concepts. The Lipid concept is a sub-concept of Small_Molecules, subsumed by the super-concept Biomolecules. Under the Lipid concept are the classes defined in the LIPIDMAPS systematic classification hierarchy. The hierarchy currently consists of 8 major lipid categories and has in total 352 lipid sub-concepts. Instances of these concepts are LIPIDMAPS systematic names of individual lipids.

The Lipid_Specification concept contains information about individual lipids and entails the following sub-concepts; Biological_Origin, Data_Specification (with a focus on high throughput data from Lipidomics), Experimental_Data (mainly mass spectrometry data values of lipids), Properties, Structural_Specification and Lipid_Identifier (that carries within it 2 other sub-concepts; Lipid_Database_ID and Lipid_Name). A Lipid instance (a systematic name) relates to individuals (equivalent to attributes/column data in a database table) from Lipid_Specification via different properties, e.g *has_Mass_Spectra_Data_Values*

Relationship with other non-lipid databases:

In addition, each Lipid instance is related to other databases via the *has_DatabaseIdentifier* property. The *has_DatabaseIdentifier* property links a lipid individual to a database identifier. This ontology is designed to capture database information from the following databases, Swisprot [9], NCBI [10], BRENDA [11], KEGG [12]. The database record identifiers from each database are considered as instances of the respective database record.

Lipid Protein Interactions:

In order to model lipid protein interactions in the ontology, we added a Protein concept. The Protein concept is a descendant of Macromolecules and Biomolecules concepts. The systematic name of a protein from the SwisProt database is modeled as an instance of the Protein concept. A lipid instance is related to a protein instance by the *Interacts_With_Protein* property.

Lipids implicated in Diseases
Information of lipids implicated in disease can also be modeled. We added a primitive concept of Diseases in the ontology. A disease name is considered as a disease instance. A lipid instance is linked to a disease instance currently derived by text mining via a *hasRole_in_Disease* property.

Modelling synonyms
Due to a lack of systematic classification, a lipid molecule can have many synonyms. In the Lipid Ontology, a lipid instance is represented by its LIPIDMAPS systematic name. Synonyms of the lipids need to be modeled into the ontology. Lipid names synonyms are IUPAC names, lipid symbols and other commonly used lipid names, both scientific and un-scientific. Figure 2 shows the conceptualization of the Lipid _Specification which describes lipid names, and lipid databases identifiers. Specifically to address lipid synonyms we introduced 3 sub-concepts, IUPAC, Broad_Lipid_Name, Exact_Lipid_Name. IUPAC is directly subsumed by Lipid_Systematic_Name whereas Broad_Lipid_Name and Exact_Lipid_Name are subconcepts of Lipid_Non_Systematic_Name. For every LIPIDMAPS_systematic name, we anticipate multiple synonyms, an IUPAC name and one or more non-systematic names. The systematic name is related to an IUPAC name via a hasIUPAC_synonym property. This property is also used to relate a non systematic name to IUPAC name. Likewise, the non systematic name and IUPAC name are related to the systematic name via a hasLIPIDMAPS_synonym property.

In our conceptualization we also define a Broad_Lipid_Name as a broad synonym that can describe several lipid molecules. This concept is related to the Lipid concept and other lipid name concepts such as IUPAC, Exact_Lipid_Name via a hasBroad_Lipid_Synonym property. This means that if a non systematic name has one or more, IUPAC names/LIPIDMAPS systematic names/LIPIDMAPS identifiers/KEGG compound identifiers/LipidBank identifiers, it is actually a broad lipid synonym. In contrast, an exact lipid name is a non-systematic name that describe exactly 1 lipid molecule.

To resolve the problem of multiple synonyms in lipid nomenclature, we assembled a list of synonyms for lipids that can be found in the LIPIDMAPS database. These synonyms came from records in the KEGG and LipidBank databases that have an equivalent record found in LIPIDMAPS database. In effect, synonyms were taken from KEGG and LipidBank databases to enrich the lipid name list from LIPIDMAPS. These synonyms were subsequently grounded to their equivalent name in LIPIDMAPS. At present, the list has 36651 unique names, that covers 10103 LIPIDMAPS systematic names, 8468 IUPAC names, 22621 non-systematic names (22494 exact lipid name + 127 broad lipid names).

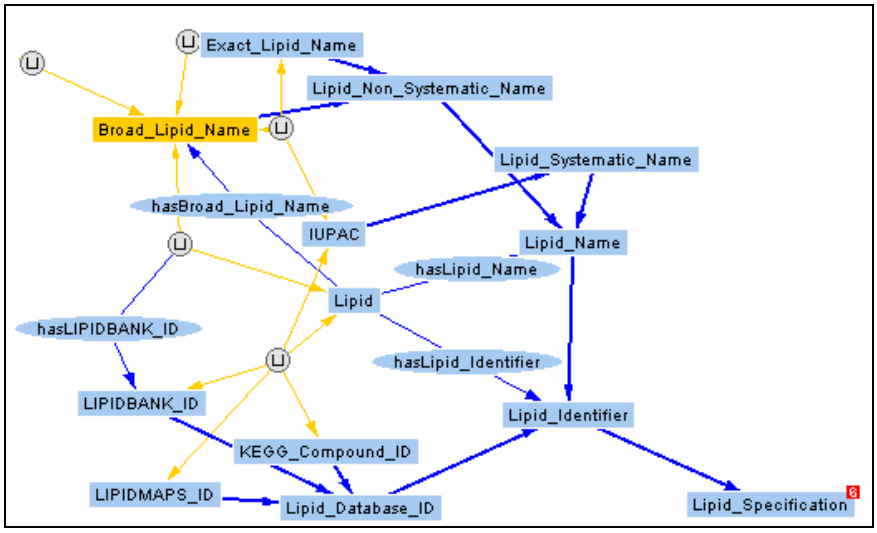

Fig. 2. Conceptualization of Lipid Specification illustrating the categorization of lipid names

Literature Specification
Of particular relevance to the application scenario, in Section 4 – is the provision of a knowledge framework where effective text mining of lipid related information can be carried out. This is supported by the Literature_Specification concept that has 10 sub-concepts, namely; Author, Document, Issue, Journal, Literature_Identifier (with a sub-concept PMID), Sentence, Title, Volume, Year. The Document concept is related to multiple concepts within the Literature_Specification hierarchy via several appropriate properties. The Document concept also has three datatype properties; author_of_Document, journal_of_Document, title_of_Document that are instantiated by author names, journal names and titles of the articles in the form of text strings. The sub-concept Sentence is related to Lipid and Protein via the property hasLipid and hasProtein. It is also related to Document via occursIn_Document property and has a datatype property, 'text_of_Sentence' that becomes instantiated by a text string from a Document found by text mining to have a lipid name and protein name or disease name occurring in the same sentence.

4 Ontology Population Workflow

In this section we describe the content acquisition; natural language processing and ontology instantiation strategy. Primarily ontology instances are generated from full texts using a text mining toolkit called the BioText Suite [13,14,15,16] which performs text processing tasks such as tokenization, part-of-speech tagging, named entity recognition, grounding, relation mining.

Content Acquisition: Our content acquisition engine takes user keywords and retrieves full-text research papers using a Pubmed search, parsing the search results and crawling the publishers' websites. Collections of research papers are converted from their original formats, e.g. pdf, to ascii text and passed to the text mining system.

Named Entity Recognition: The BioText Suite processes retrieved full-text documents and recognizes entities using a gazetteer. The gazetteer matches term lists against the token of a processed text and tags the terms found. It supports rules, e.g. for case-sensitive/case-insensitive matching, or sub/full-string matching. During gazetteer lookup, the ontology class of the term is also added as an attribute, and this is used later during the instantiation process to identify the right ontology class for population.

Separate term lists are employed for detecting lipids, proteins and diseases. The lipid name list was generated from Lipid DataWarehouse [8] containing lipid names from LIPIDMAPS, LipidBank and KEGG [12]. Each lipid name is identified by a LIPIDMAPS systematic name [17], IUPAC name, Common name and optionally other synonyms, along with a database identifier. As of April 2007, LIPIDMAPS contained 10103 entries. There were 2897 LipidBank entries and 749 KEGG entries linked to the corresponding entries in LIPIDMAPS via the database ID. All these linked entries were collapsed and grounded to their respective systematic name (explained in detail in the next paragraph). Term lists were created for each category of names: Systematic, IUPAC, broad and exact synonyms. The manually curated Protein name list from Swiss-Prot (http://au.expasy.org/sprot/) was used for grounding of proteins found in literature and further consolidated by combining all canonical names and synonyms. Grounding used the Swiss-Prot ID. A disease term list was created from the Disease Ontology of Centre for Genetic Medicine (http://diseaseontology.sourceforge.net) and used for grounding disease names.

Normalization and Grounding: Entities recognized in the previous step need to be normalized and grounded to the canonical names, before instantiation. Protein names were normalized to the canonical names entry in Swiss-Prot. The grounding is done via the Swiss-Prot ID. For lipid names, we define the LIPIDMAPS systematic name as the canonical name, and for grounding, LIPIDMAPS database ID is used. Disease names are grounded via the ULMS ID.

Relation Detection: In this step we identify the Lipid-Protein and Lipid-Disease relations, using the grounded entities. We adopt a simple relation mining approach whereby two entities are said to be related if they co-occur in a sentence. Thus, every document is parsed to extract sentences and then co-occurrence detection is invoked. To reduce false positives, we require that the sentence contain one relation keyword. All other sentences are skipped. From the resulting collection, Lipid-Protein or Lipid-Disease pairs are returned along with the respective sentences in which they co-occur. The latter could possibly be used for human validation during the knowledge retrieval step.

Ontology Population: Here we collect all the mined knowledge from the previous steps to instantiate the ontology. The grounded entities are instantiated as class instances into the respective ontology classes (as tagged by the gazetteer), and the relations detected are instantiated as Object Property instances. We wrote a custom script using the JENA API (http://jena.sourceforge.net/) for this purpose.

4.1 Population Performance Analysis

To the best of our knowledge, there is no lipidomics-related corpus for evaluating literature mining and ontology population. We are in the process of building one with biologists from the Lipidomics group at the Centre for Life Sciences, NUS, Singapore. For this paper, we provide a preliminary performance analysis of the text processing and ontology population system by assessing the complete lipid-protein interaction mining task. This started with a PubMed literature search for the query "lipid interact* protein" with our content acquisition engine that identified 495 search results for the time period July 2005 to April 2007. 262 full-text papers were successfully downloaded. The remaining papers were from journals not subscribed to by our organization or had no download-able link to the full paper.

After named entity recognition and relation detection, 121 documents in which no lipid-protein relations were detected were omitted. Ontology instantiation was carried out with the remaining 141 documents. The named entity recognition (NER) component detected 186 lipid names and 528 protein names. After normalization and grounding, there were 92 LIPIDMAPS systematic names, 52 IUPAC names, 412 exact synonyms, 6 broad synonyms and 319 protein names. Cross-links to 59 Lipidbank entries and 41 KEGG entries were also established. The brute-force co-occurrence detection yielded over 1356 sentences. After the relation word filtering, there were only 683 interaction sentences. The 92 LIPIDMAPS names were instantiated into 35 unique classes under the Lipid name hierarchy, at an average of about 2.6 lipids per class. The ontology instantiation process took 22 seconds overall. The experiments have been done on a 3.6 Ghz Xeon Linux workstation with 4 processors and 8GB RAM.

5 Knowledge Navigation for Lipidomics

The development of the ontology-centric knowledge-delivery platform results in a rich knowledge base of instantiated text segments. Typically such an OWL-DL knowledgebase is accessed through highly expressive DL-query languages that have complex syntactic query languages not suitable for domain experts [18]. nRQL is the prominent OWL-DL query language that we used which extends the existing capabilities of RACER with a series of query atoms. While some tools exist which facilitate enhanced end user operability of this query language [19, 20, 21] these implementations are of academic prototype scale and their adoption has yet to be widespread. Here we describe a new tool for the navigation of A-box instances, in our case ‘text segments’ which allows users to build graphical queries which are converted to query language syntax and issued to the reasoner.

9

5.1 Knowlegator

The Knowledge Navigator (Knowlegator) receives OWL-DL ontologies as input and passes them to RACER, after which it enters into a dialogue with RACER and issues a series of commands to query elementary features of the ontology for visual representation in the components panel. The navigator consists of three main panels, a Components panel, the Editor panel and the Output panel (Figure 3). The Components panel renders the ontology as a tree structure of concepts, roles and instances. Concepts are pre-queried to retrieve their respective number of instances and occurrences of object properties. This panel allows drag and drop functionality for query formulation. The Editor Panel is structured as a tabbed pane providing rapid switching between groups of functionalities. The 'Ask a Question' Tab contains the query canvas where questions can be formulated by dragging and dropping an element from the tree structure in the Component panel. Each dropped item is associated with an automatically formulated nRQL query. Dragging a single concept invokes the retrieval of all the individuals of a particular concept. Likewise dragging a named role (object property) queries instances specified in the domain and range of the particular role. In the query canvas a complex query built by extending simpler queries through 'right click' enabled instantiated-object property lookup. A separate window shows a query result specifically in the bottom panel the full text of a sentence is rendered. In addition to facilitating nested role queries through domain-property-range expansion the tool facilitates the identification of (instantiated) relations between any two concepts dragged to the canvas. This provides users with additional entry point to building graphical queries which can be subsequently customized. This is achieved by an exhaustive cascade of nRQL role queries to the ontology.

5.2 Lipidomics Application Domain

The intended user of the system is a researcher who specializes in lipidomics. Lipidomics is a recent research methodology that measures the composition & fluctuation of lipids at the system level of a living system in a high throughput manner. This type of user would like to ascertain the identity of lipids found in his or her experimental work and obtain all other information associated to the lipid in question. In short, they are looking for a, *one stop shop,* knowledge aggregator. Typically, for post-experiment analysis, a user has to visit multiple website or read 5-6 papers to find out the information that they want. Even then, the information that they obtain may be fragmented. Such users are typically not IT savvy and probably only proficient with a Windows environment. When such users do adopt expert or customized software for their work, they can't do without an intuitive GUI interface. Furthermore spending too much learning a new system is not considered useful even if there is a longer term benefit.

10

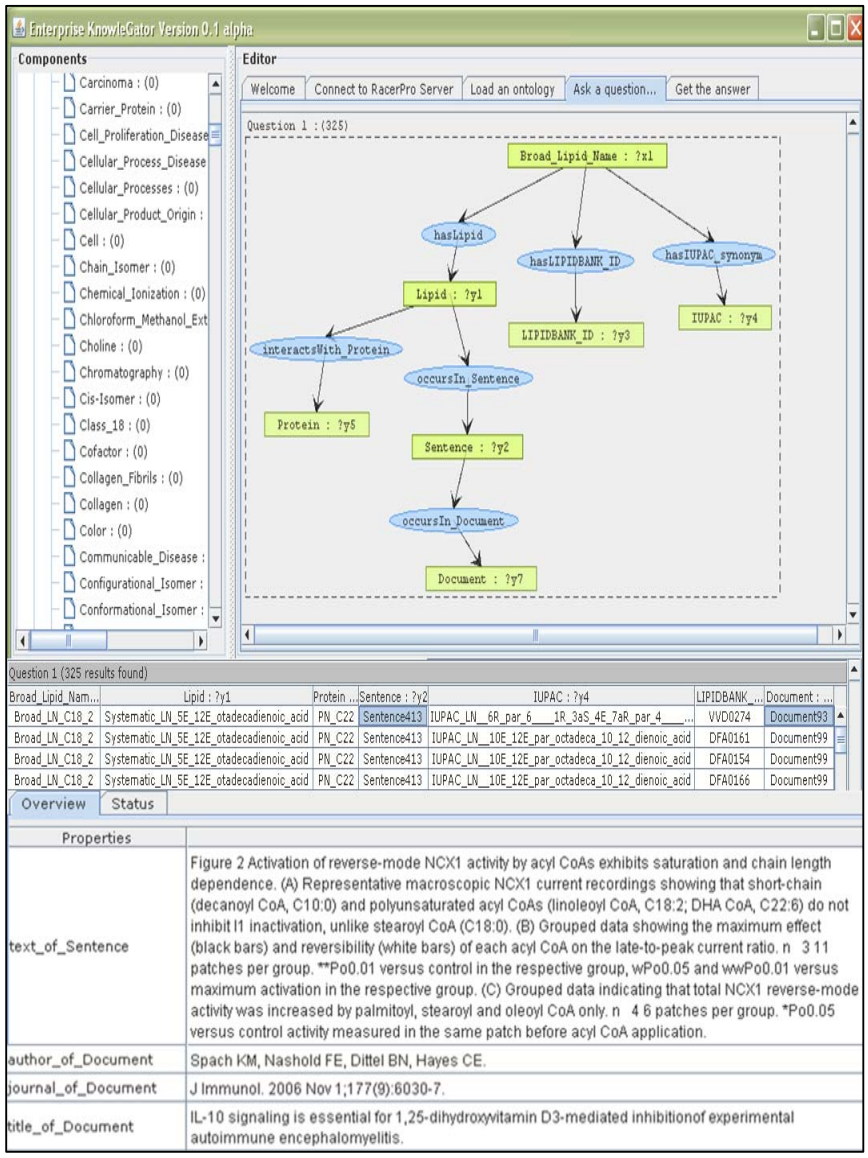

Fig. 3. Query interface of Knowlegator, showing a query for documents that contain sentences describing the interaction of proteins with lipids, and their corresponding lipid synonyms.

11

Lipidomics User Tasks:
The major knowledge-based task of a lipidomics researcher is to resolve the identity of a lipid entity to a given systematic lipid classification. The researcher can have multiple starting-points e.g. raw mass spec data, a common name from the literature or systematic name from an automated annotation pipeline, that must be translated to another classification system based on the users knowledge of lipid synonyms. Using a systematic lipid classification the user can determine or infer the possible functions / biochemical properties of the lipid. Further examination of the relationships in which a particular lipid or class of lipids participates e.g. which types of proteins a lipid interacts with, allows the researcher to make inferences regarding the metabolic process in which it participates or the role of the lipid in a cellular function or disease. Integral to these tasks is the frequent consultation with, and navigation of, the scientific literature using a variety of systematic and non-systematic lipid keywords.

Use Case Description:
The use case scenario of our system initiates with the pre-selection of collection of documents identified by an ad hoc query to a literature database or search engine and identifies relevant abstracts. The user identifies which collection of documents to review and sends them for full-text processing and the creation of a knowledgebase. The user does not require online access to the knowledgebase immediately after document selection and can wait for full text processing to complete. It is relevant to mention that major pharmaceutical corporations regularly make significant financial investments in the manual curation (3 or more months at a time) of scientific literature to generate targeted knowledge bases. This work is often outsourced to smaller companies where labour costs are cheaper. Our approach mirrors this scenario where the decision for a search and the actual navigation of the retrieved documents is decoupled into separate tasks. Once the knowledgebase is created the user has ad-hoc access to the knowledgebase using the concepts and relations provided in the query model of the ontology. The query model has rich domain specific semantics that the lipidomics user is already familiar with i.e. the systematic classification schemes of lipids. In our case the lipid ontology was built by a team (conceptualized by the lipid experts and created by ontology engineers).

5.3 Query Paradigm Comparison

Whereas searching online scientific literature databases provides sufficient ad-hoc access to abstracts it does not facilitate deep search of the full text of the documents. Systematic names of enzymes, lipids and other medical terminologies are rarely included in scientific abstracts. Additionally queries to online literature databases are limited to keyword and Boolean expressions and the traversal of literature resources is frequently based on author supplied keywords. More advanced searches of the scientific literature rely either on browsing manually curated database entries or searching the results of text mining platforms deposited in relational databases. These typically have form based web interfaces limiting the types of queries that can be issued to the database. As a result users may be required to directly interact with the relational database to pose queries that were not perceived necessary or relevant when the web portal to the database was created. This is not untypical. It is at this point where the user loses access to the knowledge resources.

For this reason we further comment on the capabilities of the ontology-centric visual query paradigm by contrasting query through the Knowlegator interface with that of a the same query made directly to a relational database with equivalent content. For example, querying for documents which contain sentences describing “lipids that interact with proteins” can be more easily formulated from the ontology by visual query than in the relational database scenario (Figures 3 and 4). Figure 3 also highlights the inclusion of Broad Lipid Names in the query such that synonyms of the lipids, in different classification schemes can be readily queried at the same time. In the database scenario, to make this query each concept should be modeled into a separate table and the relations are modeled into additional connection tables (Figure 4) to reduce redundancies. Every time there is a new relation, there must be a new relationship table. The SQL query (Figure 4) for the mentioned statement would require multiple table-joins and is not particularly intuitive to a user with no prior knowledge of the database. Using Knowlegator, the statement can be easily retrieved through a series of right mouse-clicks and selecting the required options.

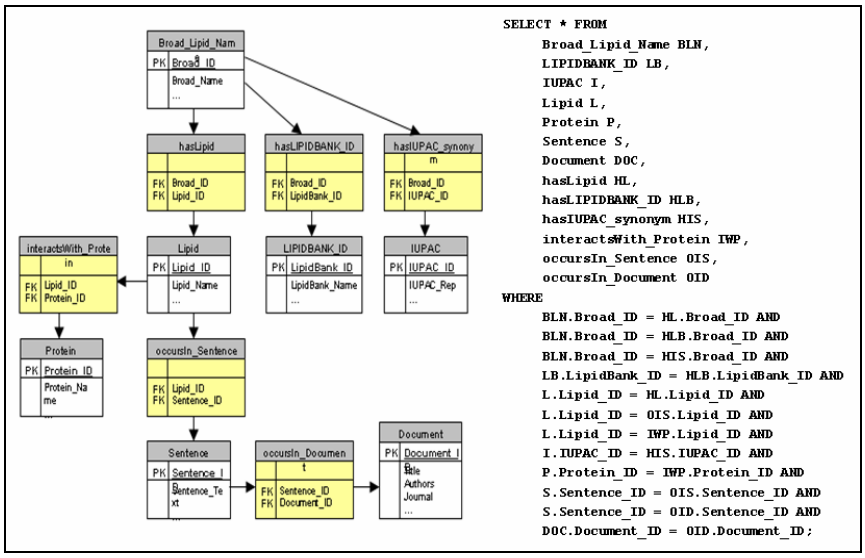

Fig. 4. A relational database query for documents that contain sentences describing the interaction of proteins with lipids and their corresponding lipid synonyms.

6 Conclusion

The challenge in our Lipidomics scenario is the navigation of large volumes of complex biological knowledge typically accessible only in legacy unstructured full-text format. This was achieved through the coordination of distributed literature sources, natural language processing, ontology development, automated ontology instantiation, visual query guided reasoning over OWL-DL A-boxes. The major innovations were to: translate the results of natural language processing to instances of a ontology domain model designed by end users; exploit the utility of A-box reasoning to facilitate knowledge discovery through the navigation of instantiated ontologies and thereby enable scientists to identify the importance of newly identified lipids through their known associations, synonyms and interactions with classes of protein and diseases.

References:

[1] Baker, C.J.O. and Cheung, K.H. (Eds.) (2006) Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, Springer.
[2] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, L. Ding, P. Kolari, A.P. Sheth, I.B. Arpinar, A. Joshi, T. Finin, Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection, 15th International World Wide Web Conference (WWW2006), Edinburgh, Scotland, UK, May 2006
[3] Witte, R., Kappler, T. and Baker, C.J.O. (2006a) 'Ontology Design for Biomedical Text Mining', In Baker, C.J.O. and Cheung, K.H. (Eds.) (2006) Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, Springer. Chap. 13, pp. 281–313.
[4] Haarslev, V., Moeller, R., Wessel, M.: Querying the semantic web with racer + nrql. In Bechhofer, S., Haarslev, V., Lutz, C., Moeller, R., eds.: CEUR Workshop Proceedings of KI-2004 Workshop on Applications of Description Logics (ADL 04), Ulm, Germany (2004)
[5] Wenk MR. The emerging field of Lipidomics. Nature Review Drug Discovery, July 2005, Vol. 4, No. 4, pp.594-610.
[6] IUPAC-IUB Commission on Biochemical Nomenclature (CBN). The nomenclature of lipids (recommendations 1976). 1977. Eur. J. Biochem. 79: 11–21; 1977. Hoppe-Seylers Z. Physiol. Chem. 358: 617–631; 1977. Lipids. 12: 455–468; 1977. Mol. Cell. Biochem. 17: 157–171; 1978. Chem. Phys. Lipids. 21: 159–173; 1978. J. Lipid Res. 19: 114–128; 1978. Biochem. J. 171: 21–35 (http://www.chem.qmul.ac.uk/iupac/lipid/).
[7] Fahy E, Subramaniam S, Brown HA, Glass CK, Merrill AH Jr, Murphy RC, Raetz CR, Russell DW, Seyama Y, Shaw W, Shimizu T, Spener F, van Meer G, VanNieuwenhze MS, White SH, Witztum JL, Dennis EA. A comprehensive classification system for lipids. Journal of Lipid Research, May 2005, Vol. 46, pp.839-862.
[8] Koh J and Wenk MR Lipid Data Warehouse (Unpublished)
[9] Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research, January 2003, Vol 31, No.1, pp.365-370.

[10] D.L. Wheeler, C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler, T.A. Tatusova, B.A. Rapp, Database resources of the national center for biotechnology information, Nucl. Acids Res. 28 (1) (2000) 10–14,
[11] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, D. Schomburg,BRENDA, the enzyme database: updates and major newdevelopments, Nucl. Acids Res. 32 (Database issue) (2004) D431–D433.
[12] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acid Research, January 2004, Vol. 32(Database issue), pp.D277-280.
[13] BioText Suite: Tools for Mining Biomedical Literature. http://research.i2r.a-star.edu.sg/kanagasa/BioText/. 2006.
[14] Doreen Tan, SL Goh, K. Rajaraman, S. Swarup, VB Bajic, Tiow Suan Sim. A user-friendly text-mining tool for streptomyces biology. Combined Scientific Meeting, Singapore, 2005.
[15] Kanagasabai Rajaraman, Zuo Li, V.B. Bajic. Extracting Transcription Factor Relations from Biomedical Texts. 5th Hugo Pacific Meeting & 6th Asia-Pacific Conference on Human Genetics, Singapore, Nov 2004.
[16] Kanagasabai Rajaraman and Ah-Hwee Tan. Mining Semantic Networks for Knowledge Discovery. IEEE Conference on Data Mining (ICDM'03), Florida, USA, pp 363-366, 2003.
[17] Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH Jr, Murphy RC, Raetz CR, Russell DW, Subramaniam S. LMSD: LIPID MAPS structure database. Nucleic Acid Research, January 2007, Vol. 35(Database issue), pp.D527-D532.
[18] Smith, B., Ceusters,W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., Rosse, C.: Relations in biomedical ontologies. Genome Biology 6 (2005)
[19] A Fadhil and V. Haarslev, GLOO: A Graphical Query Language for OWL ontologies . OWL: Experience and Directions 2006, Athens, 2006.
[20] Kosseim, L., Siblini, R., Baker, C.J.O. and Bergler, S. (2006) ‘Using Selectional Restrictions to Query an OWL Ontology’, In International Conference on Formal Ontology in Information Systems (FOIS 2006), Baltimore, Maryland, USA.
[21] Baker, C.J.O., Shaban-Nejad, A., Su, X., Haarslev, V. and Butler, G. (2006a) ‘Semantic Web Infrastructure for Fungal Enzyme Biotechnologists’, Journal of Web Semantics, Vol. 4, No. 3. Special issue on Semantic Web for the Life Sciences.