THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

A research division of the U.S. National Library of Medicine

TECHNICAL REPORT

Advanced Library Services

Developing a Biomedical Knowledge Repository to Support Advanced Information Management Applications

September 2006

Olivier Bodenreider, M.D., Ph.D.
Thomas C. Rindflesch, Ph.D.

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Table of Contents

1 Background 1

2 Project Objectives 1

3 Project Significance 1

4 Methods and Procedures 2

4.1 Overview 2

4.2 Extracting Predications from Text 2

4.2.1 SemRep 2

4.2.2 SemGen 3

4.2.3 Other systems 3

4.3 Converting Structured Data into a Common Format 4

4.3.1 Description of Resources 4

4.3.2 Representing normalized knowledge 5

4.3.3 Pilot project: Converting Entrez Gene to RDF 6

4.4 Integrating predications: Biomedical Knowledge Repository 8

4.4.1 Overview 8

4.4.2 Origin of the predications. 8

4.4.3 Metainformation associated with the predications. 8

4.4.4 Storing the predications. 9

4.4.5 Querying the predications. 9

4.4.6 Integrating predications. 9

4.4.7 Estimated size of the repository. 9

4.5 Exploiting the Repository: Semantic Medline Web Portal 10

4.5.1 Background 10

4.5.2 Implementation 10

4.5.3 Using Semantic Medline 11

5 Evaluation Plan 13

5.1.1 Evaluating extraction 13

5.1.2 Evaluating integration 14

5.1.3 Evaluating applications 14

6 Project Schedule 14

7 Project Resources 14

8 Summary 15

9 References 15

10 Appendix 21

10.1 Appendix A: Example of XML representation 21

10.2 Appendix B. Example of RDF representation 21

1 Background

Expert assessment of the nation's health care system suggests that it is slow in translating knowledge into practice and that patient care is not keeping abreast of advances in basic research [1]. The American Medical Informatics Association recently proposed a plan for improving health care delivery. The plan [2] focuses on clinical decision support, which "encompasses a variety of approaches for providing clinicians, staff, patients or other individuals with timely, relevant information that can improve decision making, prevent errors, and enhance health and health care." The National Library of Medicine (NLM) can make a substantial contribution to improving national health by making a wide range of biomedical resources readily accessible to advanced information technology.

There is a large, and growing, amount of online health-related information. Some resources are in the form of text readily accessible only to humans; examples include MEDLINE/PubMed and ClinicalTrials.gov. Other information is structured and includes biomedical vocabularies, clinical and molecular biology knowledge bases, and model organism annotation databases. Within NLM, examples include the Unified Medical Language System and Entrez Gene. There are also several specialized biomedical databases, for example the BIND database [3] and PharmGKB [4], as well as commercially curated drug databases, such as Micromedex DRUGDEX [5] and DrugDigest [6].

The growth of online information resources for biomedicine is outstripping access applications that could allow users to maximally exploit those resources. In order to take full advantage of available resources, the Library needs to provide services beyond traditional information retrieval (such as PubMed). Emerging research provides the underpinnings for developing applications that accommodate the growth in online information. Advanced applications manipulate information, not just text, and provide visualization of results and interconnections among multiple sources. Some research concentrates on extracting information from text [7-11]. Other emerging systems focus on using the information extracted; examples include automatic summarization (which pinpoints the most relevant information in large document sets) [12, 13], question answering (which provides "just in time" information)[14-17], and knowledge discovery (which uses extracted information for hypothesis generation) [18, 19].

2 Project Objectives

We propose a research initiative for accommodating applications that more effectively process online information than is currently possible. The proposal has several core objectives: a) Create a comprehensive repository of executable biomedical knowledge drawn from both the research literature and structured databases; b) Develop advanced applications that directly access the repository; c) Exploit and extend ongoing research at LHNCBC.

3 Project Significance

This project is significant from several points of view. It creates a comprehensive online resource that integrates data resources across the biomedical spectrum into a seamless repository with distributed, interoperable architecture, thus allowing NLM's extensive biomedical information resources to be effectively exploited by emerging applications in automatic information management. As key components of an infrastructure for translational research, such applications con-

tribute to enhanced understanding of the processes underpinning disease and advances in patient care. A more intangible, but nonetheless significant, aspect of this project is that it supports NLM's prominent role as a leader in affording comprehensive health information to both professionals and the public. Finally, this project provides a framework for trans-NIH collaborative projects. In actively constructing informatics resources for basic research and information dissemination, the **Biomedical Knowledge Repository (BKR)** and associated applications have the potential to play a significant role in translating scientific advances into improvements in clinical practice and public health.

4 Methods and Procedures

4.1 Overview

We discuss issues involved in constructing a **Biomedical Knowledge Repository** and illustrate an emerging Web application that exploits a preliminary version. The **SemRep** [20, 21] and **SemGen** [22, 23] natural language processing systems will be used to extract information in the form of semantic predications consisting of arguments and a predicate that represent relations between concepts asserted in text. Information in biomedical structured resources will be converted into a common format (also predications) and all information will be integrated into the repository. Finally, a significant aspect of this project is the development of applications that exploit the **Biomedical Knowledge Repository**.

4.2 Extracting Predications from Text

Two programs developed at LHNCBC (**SemRep** and **SemGen**) will initially be used to extract semantic predications from ClinicalTrials.gov narrative and MEDLINE/PubMed citations for the repository. **SemRep** was devised to apply to the clinically oriented research literature, while **SemGen** addresses the genetic etiology of disease.

4.2.1 SemRep

SemRep is a rule-based symbolic natural language processing system developed for the biomedical research literature. As the first step in identifying semantic predications, **SemRep** produces an underspecified (or shallow) syntactic analysis based on the SPECIALIST Lexicon [24] and the **MedPost** part-of-speech tagger [25]. The most important aspect of this processing is the identification of simple noun phrases. In the next step, these are mapped to concepts in the Metathesaurus using **MetaMap** [26]. Syntactic analysis in Table 1-line (2), for example, contains Metathesaurus concepts and semantic types (abbreviated) for the sentence in Table 1-line (1).

Sentence	Phenytoin induced gingival hyperplasia	(1)
Syntactic analysis	[[head(noun(phenytoin)), metaconc('Phenytoin':[orch,phsu]))], [verb(induced)], [head(noun(['gingival hyperplasia')), metaconc('Gingival Hyperplasia':[dsyn]))]]	(2)
Semantic Network	'Pharmacological Substance' CAUSES 'Disease or Syndrome'	(3)
Semantic Predication	Phenytoin CAUSES Gingival Hyperplasia	(4)

Table 1. SemRep analysis

SemRep relies on structures such as that in Table 1-line (2) to translate syntactic structures into semantic predications. One aspect of this processing is "indicator" rules which map syntactic elements (such as verbs and nominalizations) to predicates in the Semantic Network, such as TREATS, CAUSES, and LOCATION_OF. Argument identification rules (which take into account coordination, relativization, and negation) then find syntactically allowable noun phrases to serve as arguments for indicators. If an indicator and the noun phrases serving as its syntactic arguments can be interpreted as a semantic predication, the following condition must be met: The semantic types of the Metathesaurus concepts for the noun phrases must match the semantic types serving as arguments of the indicated predicate in the Semantic Network. For example, in the structure given in Table 1-line (2) the indicator *induced* maps to the Semantic Network relation in Table 1-line (3).

The concepts corresponding to the noun phrases *phenytoin* and *gingival hyperplasia* can serve as arguments because their semantic types ('Pharmacological Substance' (phsu) and 'Disease or Syndrome' (dsyn)) match those in the Semantic Network relation. In the semantic predication produced as output (Table 1-line (4)), the Metathesaurus concepts from the noun phrases are substituted for the semantic types in the Semantic Network relation.

4.2.2 SemGen

SemGen was adapted from **SemRep** in order to identify semantic predications on the genetic etiology of disease. The main consideration in creating **SemGen** was the identification of gene and protein names as well as related genomic phenomena. For this **SemGen** relies on **ABGene** [27], in addition to **MetaMap** and the Metathesaurus. Since the UMLS Semantic Network does not cover molecular genetics, ontological semantic relations for this domain were created for **SemGen**. The allowable relations were defined in two classes: gene-disease interactions (ASSOCIATED_WITH, PREDISPOSE, and CAUSE) and gene-gene interactions (INHIBIT, STIMULATE, and INTERACTS_WITH).

4.2.3 Other systems

The information submitted to the repository by **SemRep** and **SemGen** could be supplemented with output from other natural language processing technologies that produce relationships. Phenotypic information from clinical narrative could be made accessible with the NLP system described by Friedman et al. [28]. For molecular biology phenomena, several systems use syntactic templates and shallow parsing to produce a variety of relations [29], including gene and protein functions [30], protein interactions [7], and protein modifications such as phosphorylation [31].

Friedman et al. [32] use extensive linguistic processing for relations on molecular pathways, while Lussier et al. [9] use a similar approach to identify phenotypic context for genetic phenomena.

4.3 Converting Structured Data into a Common Format

Since relations in structured resources are already represented in some kind of formalism, their conversion to a common format is somewhat easier than extraction from text. However, the challenges in normalizing such knowledge are not unlike those encountered with textual data. In both cases, knowledge extraction involves syntactic issues (i.e., the formalism used to represent knowledge) and semantic issues (i.e., the meaning of the terms used to represent data and their interrelations). Various formalisms are used to represent structured data, including relational databases, tables (e.g., Excel spreadsheets), graphs, etc. Currently, no universal conversion mechanism is available. Moreover, the semantics of the data is often implicit (especially for relationships) or limited, for example, to short column names in a database schema. The objective of the conversion is the creation of "normalized knowledge". This effort, in the context of knowledge management, is somewhat equivalent to term normalization (i.e., the models of lexical resemblance used, for example, to identify candidate synonyms in the UMLS [24]). Knowledge normalization ensures that the same entity referred to in different resources is ultimately identified in such a way that it is recognized as a unique thing. Knowledge normalization forms the basis for information and data integration.

4.3.1 Description of Resources

Over the past twenty years, NLM has developed many knowledge resources, from various perspectives. Terminological resources such as the Unified Medical Language System (UMLS) result from the integration of many existing terminologies and ontologies, represented in a common formalism – UMLS' Rich Release Format – and partially normalized. While synonymous terms are grouped together as names for a given concept, synonymous relationships are typically not identified as such [33]. Nevertheless, with some 8 million relations, the UMLS Metathesaurus constitutes an important resource, providing mostly hierarchical relations (useful as a "backbone" for the **Biomedical Knowledge Repository**) and co-occurrence relations.

Another source of structured knowledge is represented by the many databases available under the umbrella of the NCBI's Entrez system. While some databases such as MEDLINE/PubMed and OMIM (Online Mendelian Inheritance in Man) are mostly textual resources, many other NCBI databases contain predominantly structured information. Entrez Gene, for example, is a gene-centric resource in which a record provides gene properties such as names, associated diseases, function, sequence, etc [34]. While numerous links have been created across the resources in the Entrez system for navigation purposes (e.g., between Gene and OMIM), such links typically require human interpretation and therefore cannot be used for knowledge discovery purposes in high-throughput systems.

Finally, structured databases and knowledge bases are also available outside NLM. One example of large, publicly available genomic resource is represented by the genome annotations for the major model organisms (fly, mouse, yeast, worm, etc). Here, the existence of a controlled vocabulary – the Gene Ontology – has contributed to developing unified functional annotations, integrated in systems such as the Mouse Genome Database [35] and the Saccharomyces Genome

Database [36]. In the clinical domain, the various knowledge bases of drug information constitute important resources about drug-drug interactions, drug kinetics and metabolism, and indications and contraindications. DRUGDEX [5], produced by Micromedex, is an example of such a resource.

4.3.2 Representing normalized knowledge

One of the strengths of the World Wide Web is that it relies on textual information, easily created and interpreted by humans. This is also one of its principal limitations. The absence of explicit semantics prevents agents from being able to make sense of the information on the Web. This is the motivation of the Semantic Web [37], which aims at creating a vast collections of integrated and interoperable resources. There is an obvious parallel between the Semantic Web and the **Biomedical Knowledge Repository** we propose to create. The technologies developed for the Semantic Web offer possible solutions to some of the challenges we face, including selecting a formalism for normalized biomedical knowledge and identification issues for biomedical entities [38]. It is worth noting that some of these issues are still actively being debated in the Semantic Web community, especially in the Semantic Web Health Care and Life Sciences Interest Group [39].

Formalism. Many of the resources produced by NLM and other organizations are available in XML, the eXtensible Markup Language. However, the semantics of XML is limited. In addition to XML, the World Wide Web Consortium (W3C) has produced the specifications of other formalisms for representing resources (RDF/S, the Resource Description Framework) and ontologies (OWL, the Web Ontology Language). Collectively know as Semantic Web technologies, these specifications define the building blocks of the Semantic Web [40].

Of particular interest to us is the Resource Description Framework. RDF extends the capabilities of the extensible markup language XML as it enables many-to-many relationships between resources and data. The resulting structure is a graph in which the nodes are resources (identified by a Uniform Resource Identifier or URI) or data (e.g., strings, numerals) and the edges are relationships (called properties). The basic unit in RDF is therefore the equivalent of a (concept, relationship, concept) triple, similar to a relation in the UMLS Metathesaurus or a predication extracted by **SemRep**. RDF integrates limited inference rules, enabling, for example, the definition of subclasses and subproperties.

Some extensive resources such as UniProt [41] have already been converted to RDF [42]. The BioRDF [43] task force of the W3C Semantic Web Health Care and Life Sciences Interest Group currently investigates methods whereby existing biomedical resources can be converted to RDF. Such methods include XSLT, GRDDL and DB2RDF, among others. The eXtensible Stylesheet Language Transformation (XSLT) [44] uses a stylesheet approach to converting XML to RDF. GRDDL (Gleaning Resource Descriptions from Dialects of Languages) [45] specifies associations between markup languages – including XLM – and RDF. Finally, DB2RDF is used to convert databases to RDF.

Identification issues. In order for RDF triples to form a graph – and for integrated knowledge to be interoperable – entities (i.e., the nodes in the RDF graph) and relationships (i.e., the edges in the graph) must be identified consistently and unambiguously. For example, if the disease *Neu-*

rofibromatosis 2 is identified by the code *SNOMEDCT:92503002* in one resource (annotated with SNOMED CT) and by the code *MESH:D016518* in another (annotated with the Medical Subject Headings), the RDF triples involving *SNOMEDCT:92503002* and *MESH:D016518* will not come together as expected unless both resource are converted to the other annotation system or mappings are created between the two systems. For example, the UMLS Metathesaurus could be used to convert or bridge between MeSH and SNOMED CT, in this case through the concept *C0027832*. The predications extracted from the literature by **SemRep** already use UMLS codes to identify biomedical entities.

From a technical perspective, several technologies have been developed by various communities to implement identification mechanisms for RDF. There are three major identification mechanisms:

- LSID (Life Science Identifier), promoted by the Life Sciences community [46]. Examples of applications using LSID include Taverna [47] and resources created by the BioPathways Consortium [48].
- Solutions based on the HTTP protocol (Unified Resource Identifiers (URIs), Names (URNs) and Locators (URLs)), promoted by the W3C [49].
- ARK (Archive Resource Key), promoted by the Digital Library community [50].

There are important differences among three mechanisms regarding location independence, backward compatibility, resolution mechanism and versioning. It is unclear at this time what mechanism would suit our needs best.

4.3.3 Pilot project: Converting Entrez Gene to RDF

As a proof of concept, we converted the Entrez Gene database into RDF [51]. The entire Gene database in its native ASN.1 format was downloaded by FTP from the NCBI website (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/) and later converted to XML using the program **gene2xml** provided by NCBI. Using the eXtensible Stylesheet Language Transformation (XSLT) approach [44], we mapped the element tags of the XML representation to more intuitive relationship names manually, and used them during the automatic conversion to RDF. Finally, we stored this RDF version of Entrez Gene in the Oracle 10g relational database management system, which provides support for storing and querying native RDF data. The conversion process is illustrated in Figure 1. Examples of XML and RDF representations for a Gene record are provided in appendix A and B.

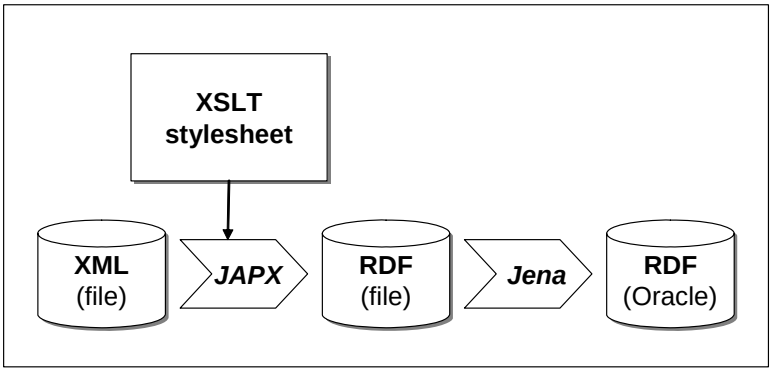

Figure 1. Overview of the conversion of Entrez Gene from XML to RDF

Mapping XML element tags to RDF properties. While XML data can be mechanically converted to RDF, the resulting RDF graph would be of limited interest because the semantics of the properties (relationships) is most often implicit in XML and needs to be made explicit. Starting from one typical Entrez Gene record, we identified all its XML element tags, corresponding to the properties of the gene, as described in the Gene record. Examples of such properties include <*Genetrack_geneid*> and <*Genetrack_update-date*>, indicating the identifier and the update date of the gene record, respectively. A total of 124 unique elements tags were identified. Because XML element tags represent gene properties, they were transformed into predicates in RDF (also called properties in RDF parlance). For example, <*Genetrack_geneid*> becomes *has_unique_geneid* and relates the gene record (subject) to its unique identifier in the Gene database (object). Note that the RDF property now conveys the notion – implicit in the XML representation – that the identifier is a unique identifier, corresponding to a primary key in a relational database. Converting XML elements to RDF properties requires familiarity with the structure and content of the original record structure and is a key component to the mapping process. After eliminating redundant and superfluous XML element tags, 106 unique RDF properties were created. The mapping of Entrez gene XML element tags to RDF properties is specified formally using the XPath language [52] and constitutes a stylesheet. This stylesheet is specific to the conversion of the Entrez Gene database.

Converting XML to RDF through XSLT. Once the stylesheet is created, it can serve as an auxiliary file for existing programs realizing the XML to RDF conversion. In other words, the major interest of this approach is that no specific code is required for the conversion, because the transformation logic resides entirely in the stylesheet. We used JAXP, the Java Application Programming Interface (API) to XML, to implement the conversion. The resulting 411 million RDF triples were then loaded in a store using Oracle 10g using the Jena API.

Lessons learned, issues and challenges. This experiment confirmed the feasibility of converting a large resource from XML to RDF. It also showed that the issues are not technical, but rather lie in the necessity of making explicit the relationships represented implicitly by element tags in XML. This step requires the manual intervention of a domain expert. In our experience, it took less than a week for one person familiar with both bioinformatics and stylesheets to formalize the mapping between XML element tags and RDF properties. Although the stylesheet created is specific to the Entrez Gene database, it is expected that part of the expertise acquired during this transformation can be applied to transforming other NCBI resources.

The major unresolved issue concerns entity identification. If the RDF graph resulting from the conversion of Entrez Gene is to be integrated with clinical or bibliographic information, the diseases associated with genes must be represented not as literals (strings) as they currently are, but by their identifiers in the corresponding clinical (e.g., SNOMED CT) and bibliographic (e.g., MeSH) controlled vocabularies, or with the concept unique identifier (CUI) in the UMLS Metathesaurus. Only after such mapping will the RDF graph integrating Entrez Gene and, say, the Metathesaurus, support queries such as *Find all genes associated with neurodegenerative diseases*. The knowledge required in this cases comes in part from Entrez gene (e.g., APP associated with Alzheimer disease and PARK3 associated with Parkinson's disease) and from UMLS

(e.g., Alzheimer's disease isa Neurodegenerative disease, Parkinson's disease isa Neurodegenerative disease).

4.4 Integrating predications: Biomedical Knowledge Repository

4.4.1 Overview

The **Biomedical Knowledge Repository** can be understood as a specialized version of the Semantic Web. It consists of an extensive collection of predications (i.e. concept-relationship-concept triples), represented in a common format, processable by computers. Biomedical terminologies and ontologies provide the concepts involved in the facts. Logical reasoners extend the capabilities of the repository by inferring new knowledge [53]. Each fact in the repository is annotated with metainformation regarding its origin (e.g. source, extraction method, timestamp, etc.), making it possible for applications exploiting the repository to select facts of interest in a given context and for a particular task. Once it is fully populated, the **Biomedical Knowledge Repository** is expected to comprise several hundred million facts, collected from hundreds of sources.

4.4.2 Origin of the predications.

The **Biomedical Knowledge Repository** (**BKR**) comprises predications form three major sources: extracted from the biomedical literature by NLP programs such as **SemRep**, converted from existing structured knowledge bases, and contributed by users (collaborative development). In our vision of the **BKR**, NLM not only contributes to populating the **BKR**, but also makes it available to the community as a framework for researchers to deposit their predications. Predications resulting from experiments, from alternative processing of the literature, and inferred from other predictions, for example, can be contributed by members of the community and made available to others. One measure of success of the **BKR** would be for it to become a standard repository for the knowledge created through biomedical research experiments. In order to maintain the consistency and integrity of the **BKR**, its developers would have to follow guidelines regarding formalism of the predications and identification mechanism for biomedical entities.

4.4.3 Metainformation associated with the predications.

Just as few users need all the vocabularies in the UMLS Metathesaurus for a given purpose, it is unlikely that all predications will be equally useful for a given task, such as question answering. Instead, a mechanism supporting the selection of relevant sets of predications is needed. The metainformation associated with each predication enables this selection and may include the source of the predication (e.g., biomedical article, database name), the method of extraction (e.g., **SemRep**, XSLT), the date of extraction, in addition to the usual metadata associated with MEDLINE/PubMed citations for predications extracted from the literature (e.g., authors, journal, MeSH main headings and checktags, etc). Researchers contributing predications to the **BKR** will be requested to annotate them with similar metainformation. Another form of contribution to the repository is to provide not predications, but annotations to existing predications. Such a contribution represents a form of collaborative curation of the repository by the community, similar to the framework developed by the SWAN project for the Alzheimer research community [54].

4.4.4 Storing the predications.

Several technological solutions, called RDF stores, have been developed for storing information in RDF format, generally implemented on top of some storage system (relational database, in-memory, file system, etc). One such open source RDF database is Sesame [55]. More recently, traditional database management systems have started offering direct support for native RDF triples. For example, version 10g of the Oracle system supports RDF in addition to the relational model. Since we need to store large amounts of both RDF (predications) and traditional information (annotations), we started experimenting with Oracle 10g while converting Entrez Gene to RDF last summer. Preliminary results are encouraging although we might face optimization issues.

4.4.5 Querying the predications.

Analogous to SQL for relational databases, SPARQL is the language for querying RDF [56]. Like SQL, SPAQRL queries have a SELECT and a WHERE clause. However, in a SPARQL query, the WHERE clause follows the pattern of a RDF triple in which at least one element of the triple is replaced by a variable. In the **BKR**, the predications to be queried would first be selected based on their annotations (metainformation). For example, a typical query supporting multidocument summarization would be as follows. Select all predications from MEDLINE/PubMed returned by a PubMed query on *metabolic syndrome*, restricted to citations from *JAMA*, *Am J Cardiol.* and *J Hypertens.*, published between 2004 and 2005. Additionally, select those predications from the UMLS Metathesaurus and Entrez Gene having at least one node in common with those from the literature. The resulting graph is expected to provide a richer, more detailed summary than would the predications from one source only.

4.4.6 Integrating predications.

The query example presented above illustrates the interest of integrating knowledge under a unique framework. First, there is a unique namespace (i.e., universe of reference) for all knowledge in the **BKR**. As mentioned earlier, we will largely rely on the UMLS for identifying biomedical entities. Second, once integrated into a graph, the predications can serve as a basis for inferencing, creating additional knowledge along the way. This represents an advantage over traditional database and information retrieval approaches. Finally, complex rules can be written to implement additional reasoning. For example, it is possible to restrict queries to those redundant predications asserted in more than 2 sources and for which the frequency of occurrence is above a certain threshold.

4.4.7 Estimated size of the repository.

Once fully instantiated, the **Biomedical Knowledge Repository** is expected to comprise several hundred million predications extracted from the literature, terminological resources and structured knowledge bases. Assuming an average of ten predications is extracted from the title and abstract of each MEDLINE/PubMed citation, we can expect about 150 million predications from MEDLINE/PubMed. (With the availability of full text articles, a significant increase is to be expected in the average number of predications extracted). The UMLS Metathesaurus records about 9 million relations, either symbolic (e.g., Parkinson disease isa Neurodegenerative disease) or statistical (e.g., co-occurrence between Parkinson disease and Dopamine agonists). Some seven million functional annotations for the major model organism databases are recorded in the Gene Ontology database. Finally, our conversion experiment with Entrez Gene yielded over 400

million RDF triples. In this experiment, the objective was to systematically represent in RDF all the information present in the original XML file. In fact, part of this information would become metainformation in the context of the **BKR** (e.g., update date of the record), resulting in a significantly smaller number of triples to be actually contributed to the repository.

4.5 Exploiting the Repository: Semantic Medline Web Portal

As a pilot project to exploit a preliminary version of the **Biomedical Knowledge Repository**, we are developing a Web portal, called **Semantic Medline**, for managing the results of PubMed searches. The portal is designed as a Java-based Web application that seamlessly integrates PubMed, **SemRep** processing of the results of the search, automatic summarization [12, 57], and, finally, visualization of the results, with links to the underlying citations and relevant additional knowledge in the UMLS Metathesaurus, the Genetics Home Reference, and Entrez Gene. Real-time access is achieved by pre-processing text from MEDLINE/PubMed abstracts and other sources with **SemRep**/**SemGen** and storing the results in a database.

4.5.1 Background

Several recent systems visualize the results of information identified in text as a way of providing users with enhanced access to the information retrieved [58]. Results are often represented as a graph of interrelated relationships [59]. The Telemakus project [60] is based on relationships identified by hand and is meant to enable knowledge discovery through interactive visual maps of linked concepts among documents. Jensen et al. [61] constructed literature networks of genes found relevant in gene expression data analysis. Analysis is based on co-occurrence of genes in MEDLINE/PubMed abstracts. Van der Eijk et al. [62] use various relations for literature-based discovery. The relationships represent co-occurrence of MeSH headings associated with MEDLINE/PubMed citations by a mapping program (MeSH terms assigned by indexers are not used). Feldman et al. [63] represent several gene-related relations (e.g. gene-gene binding; gene-phenotype; gene-disease) in a graph. The relations were extracted with a type of underspecified natural language processing. Finally, Tao et al. [64] visualize genomic information across both structured and textual databases.

4.5.2 Implementation

Semantic Medline is implemented as a three-tier, Java EE-based Web application (Figure 2). The three-tier architecture allows for the separation of user interface, application logic and data storage, providing improved performance, easier maintenance and scalability. We leverage mature open-source technologies in the development to the extent possible. The prototype runs in a Tomcat servlet container on an Apache Web server. It has been developed using the Apache Struts Web application framework. This framework encourages the use of MVC (Model-View-Controller) paradigm to provide a clean separation of application model, navigational code, and page design code through the use of Java Servlet API.

The controller is a Java servlet that mediates the application flow; the model comprises Java classes that represent the functionality of the semantic tools and the view is JSP pages that contain dynamic content. A MySQL database is used to store **Semantic Medline** data, which includes semantic predications extracted from MEDLINE/PubMed abstracts and ClinicalTrials.gov clinical study texts as well as a subset of UMLS Metathesaurus data. The database tables are pre-populated from plain text files that contain **SemRep**/**SemGen** output and Metathesaurus data us-

ing Perl scripts. Hibernate object/relational mapping (ORM) tool is used to programmatically access the database. We use such Hibernate features as database connection pooling and query caching for increased performance.

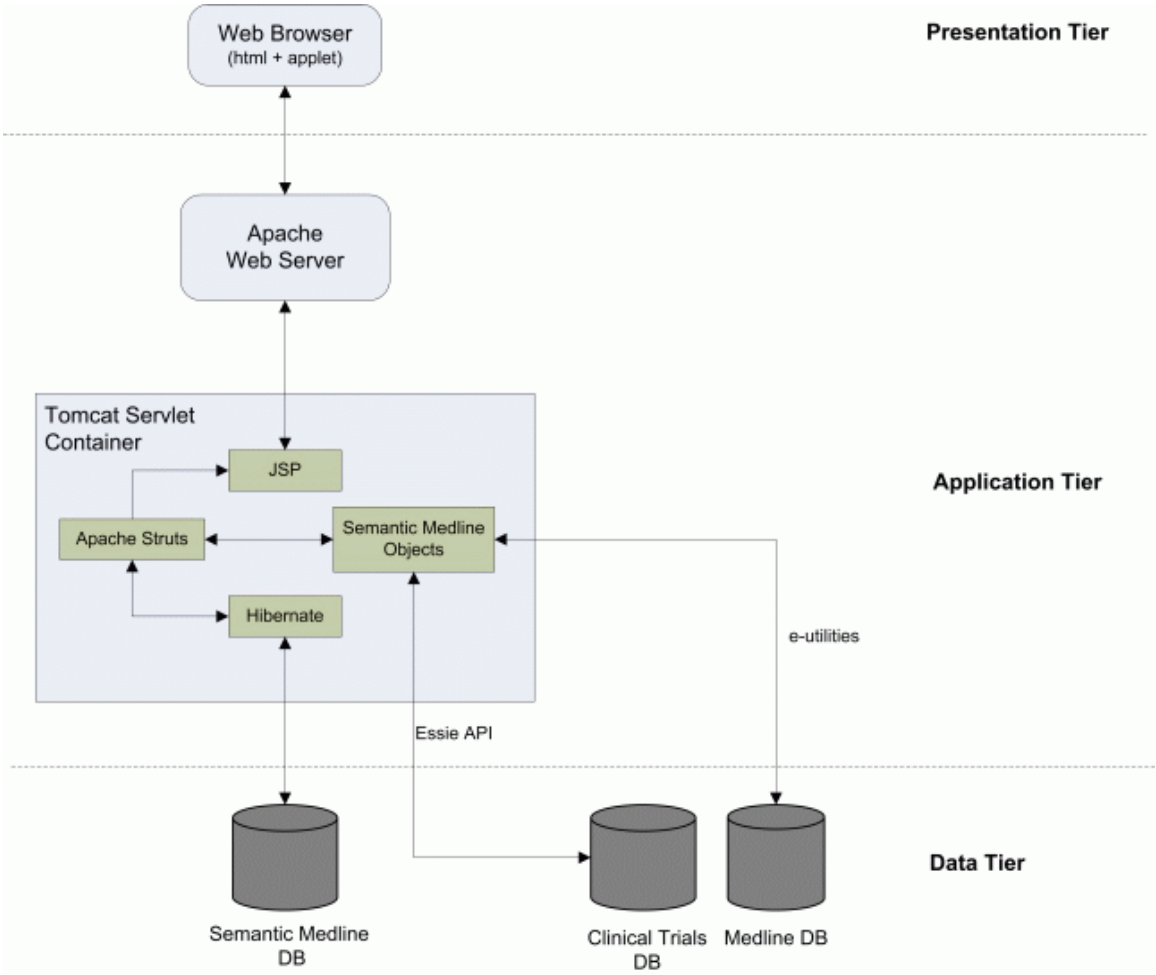

Figure 2. Semantic Medline system architecture

4.5.3 Using Semantic Medline

MEDLINE/PubMed contains more than 16 million citations (dating from the 1960's to the present) drawn from nearly 4,600 journals with biomedical relevance. Searches often retrieve large numbers of items. For example, the query "diabetes" returns 207,997 citations. Although users can restrict searches by language, date and publication type (as well as specific journals), results can still be large. For example, a query for treatment (only) for diabetes, limited to articles published in 2003 and having an abstract in English finds 3,621 items; limiting this further to articles describing clinical trials still returns 390 citations. It is difficult for a user to effectively exploit the information in this many citations. **Semantic Medline** addresses this difficulty by allowing users to summarize the results of searches focused on one of several points of view, including diagnosis and treatment of disorders, drug interactions and adverse events, genetic basis of disease, and pharmacogenomics.

A clinical scenario based on a summary focused on drug interactions illustrates the potential use of **Semantic Medline** for taking advantage of the research literature in clinical practice. In a hypothetical situation, a patient presents with peptic ulcer and tests positive for *Helicobacter pylori*. The clinician has tried several standard regimens including two different triple regimens: (proton pump inhibitor, amoxicillin, and metronidazole) and (ranitidine bismuth citrate, clarithromycin, amoxicillin); however, the patient still tests positive for *H. pylori*. (It is known that treatment of several upper gastrointestinal disorders such as peptic ulcer requires eradication of *H. pylori* [65].)

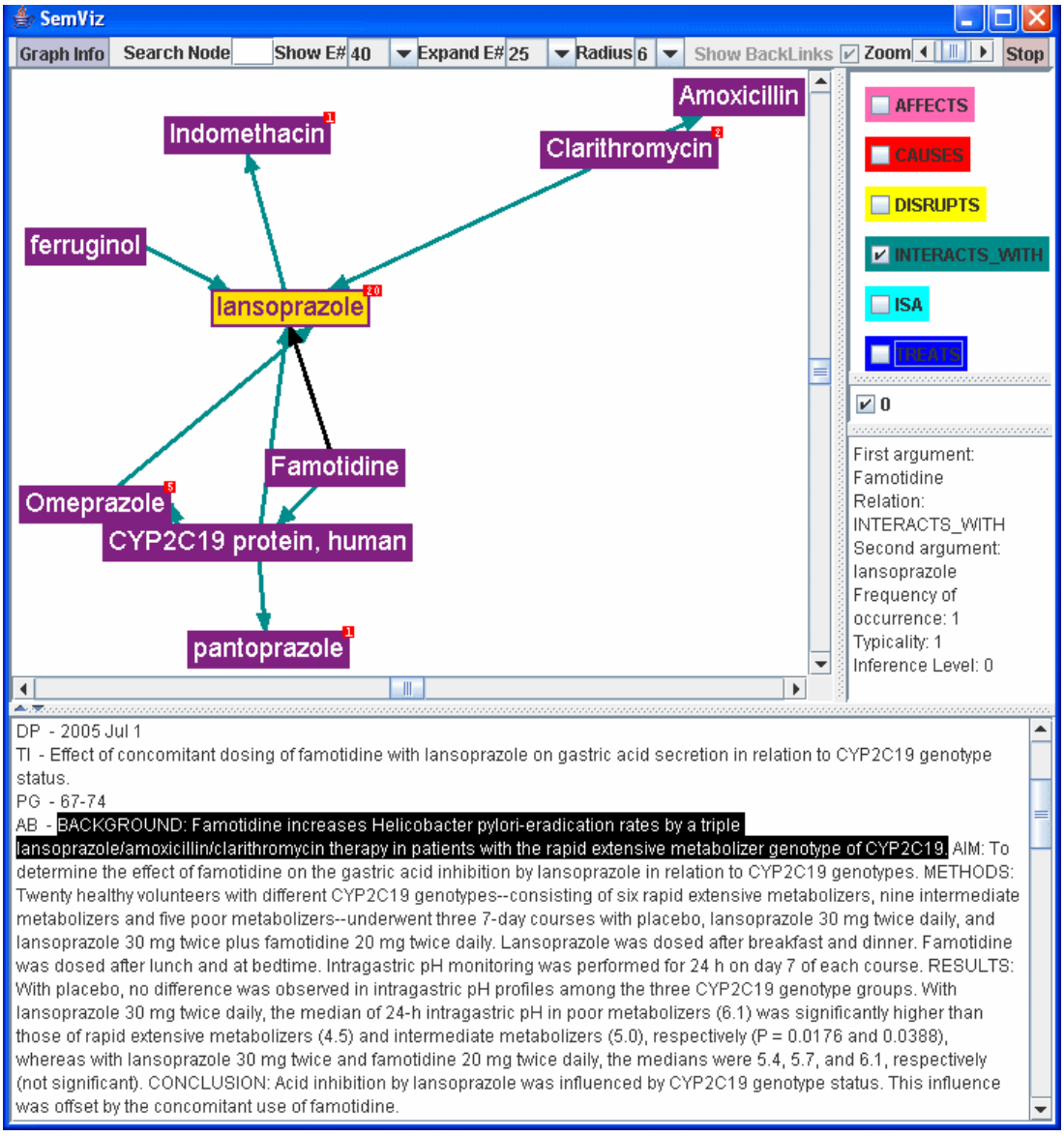

Figure 3. Summary of 564 MEDLINE/PubMed citations for lansoprazole, showing INTERACTS_WITH

In order to gain insight from recent research, the physician uses **Semantic Medline** to issue a PubMed search (limited to the past 5 years) for lansoprazole, (a proton pump inhibitor). **SemRep** extracts 5,971 predications from 564 citations returned by this search, and automatic summarization for drug interactions involving lansoprazole condenses the final list of predications to 182, which are visualized by the system as the interactive graph in Figure 3.

Figure 3 provides an informative overview of recent research on lansoprazole represented as predications asserting INTERACTS_WITH. Each predication is linked to the MEDLINE/PubMed citation that generated it. Of interest for this case are two predications in the graph: "Lansoprazole INTERACTS_WITH Famotidine" and "Famotidine INTERACTS_WITH CYPC19 Enzyme." **SemRep** extracted the first predication from a citation (15963082) with title seen in the first line of Table 2 and the second from a citation (15710002) with in the second line.

PMID	Title
15963082	Effect of concomitant dosing of famotidine with lansoprazole on gastric acid secretion in relation to CYP2C19 genotype status
15710002	Concomitant dosing of famotidine with a triple therapy increases the cure rates of Helicobacter pylori infections in patients with the homozygous extensive metabolizer genotype of CYP2C19

Table 2. MEDLINE/PubMed citations

Both citations discuss the salutary effect of combining famotidine (a histamine blocker) with lansoprazole. In the second citation, which reports on a randomized controlled trial, the authors realized between 63% to 100% eradication of *H. pylori* (depending on CYP2C19 status) with the addition of famotidine to a standard triple therapy (lansoprazole, clarithromycin and amoxicillin) and conclude that this is a promising option for patients who phenotypically are extensive metabolizers. If the patient in this hypothetical scenario falls under this category, it would be a potentially valuable regimen to pursue.

5 Evaluation Plan

We intend to follow a multifaceted evaluation plan, focusing evaluation activities on specific aspects of the project.

5.1.1 Evaluating extraction

For this project, effectiveness of information extracted from text crucially depends on the accuracy of **SemRep** and **SemGen**. We have performed linguistic evaluation across a wide variety of predicates that includes the clinical (**SemRep**) as well as the genetic (**SemGen**) domain. Among the predicates evaluated were TREATS, PREVENTS, LOCATION_OF, CAUSES, ISA, INTERACTS_WITH, AFFECTS, DISRUPTS, PROCESS_OF, PART_OF, MANIFESTATION_OF, ASSOCIATED_WITH, INHIBITS, STIMULATES, and PREDISPOSES. Recall on the extraction of semantic propositions with these predicates has ranged from 41% to 74% and precision from 68% to 84%. Previous evaluations have been reported in [12, 21-23, 57], and we will continue this evaluation paradigm.

The framework for evaluating knowledge extraction from terminological and structured knowledge bases is not as well established as the evaluation of relation extraction from text. However, several aspects are particularly important. The rules used for the conversion (e.g., the XSLT style sheet) need to be reviewed independently by several experts for a given knowledge base and integrated across databases in order to ensure consistency of the extraction of the relationships. Key to the consistency of the graph is also the mapping of literals (e.g., disease names) to concepts – the nodes in the RDF graph and their consistent representation by identifiers from authoritative sources.

5.1.2 Evaluating integration

Ideally, the graph resulting from integrating knowledge from several sources is consistent both structurally (e.g., a directed acyclic graph) and semantically (e.g., no contradicting predications, directly or inferred). Therefore structural techniques based on graph theory and semantic approaches similar to reasoning services created for description logics [66] are expected to play a central role in evaluating the quality of knowledge integration in the repository. However, the UMLS Metathesaurus, while resulting from the integration of a much smaller body of terminological knowledge, is already not always structurally or semantically consistent [67-69]. Like the Metathesaurus, the **Biomedical Knowledge Repository** will allow contradictory predications to be represented as long as they occur in the original information sources. However, consistency is expected to be found at both structural and semantic levels on limited subsets of the repository (e.g., predications extracted from the literature on a given topic published over a limited period of time).

5.1.3 Evaluating applications

When evaluating applications such as automatic summarization, it is useful to compare results against curated resources. Standard measures of performance (recall and precision) can be calculated with respect to the reference standard chosen. For treatment of disease, the British Medical Journal clinical evidence concise [70] is one (semistructured) alternative. For drug interactions and adverse events, Micromedex DRUGDEX [5], DrugDigest [6], and the First DataBank's National Drug Data File [71] can be used. A more ambitious method, requiring human experts, is task-based [72] evaluation. Ultimately, user-centered evaluations such as the one described by McKeown [73] must be considered.

6 Project Schedule

- Year 1: Pilot study, processing all of MEDLINE/PubMed, integrating Entrez Gene
- Year 2: Integrating UMLS and other NCBI resources
- Year 3: Annotating predications, opening the **BKR** to collaborative development and use
- Year 4: Fully integrated ALS system available

7 Project Resources

This project relies on a range of existing competencies and resources and has the potential to federate research energies throughout the Lister Hill Center. In addition to the efforts of the two core component projects, Semantic Knowledge Representation and Medical Ontology Research, the success of this project relies on collaboration with the Indexing Initiative, the Lexical Systems Project, and ClinicalTrials.gov. The effectiveness of applications drawing on the repository could be strengthened by providing links to Visible Human images where appropriate.

In order to accomplish the goals of this project additional staff and equipment are required: staff for creating and populating the repository and servers for providing effective access to the repository. Initially, PubMed and ClinicalTrials.gov will be processed by **SemRep** and **SemGen** and the predications extracted will be stored in the repository. Dedicated servers are required to process PubMed (more than 16 million citations) for this project.

8 Summary

NLM can make a significant contribution to improving national health by making a wide range of biomedical resources readily accessible to advanced information technology. As part of this contribution, we propose the Advanced Library Services project. The core goal of this project is to create a comprehensive repository of executable biomedical knowledge drawn from both the research literature and structured databases. Further, we are developing advanced applications that directly access the repository; such tools are seen as a vital component of the infrastructure for supporting translational research.

The **Biomedical Knowledge Repository** (**BKR**) represents a step forward compared to the individual information sources routinely queried by biomedical researchers. Knowledge in the repository is normalized, represented in a common format (i.e. concept-relationship-concept triples) and using identifiers from authoritative sources, and thus made compatible with the recommendations of the Semantic Web. Moreover, knowledge from sources including the biomedical literature, terminological resources and knowledge bases is integrated, making it possible for researchers to query heterogeneous resources seamlessly. Normalized and integrated, the knowledge available in the **BKR** can be processed by humans, but is also accessible to agents, supporting data mining and knowledge discovery applications. Finally, the applications exploiting the **BKR** can take advantage of metainformation stored with the knowledge and select various subsets of it according to the task at hand.

As a pilot project to exploit a preliminary version of the **Biomedical Knowledge Repository**, we are developing a Web portal, called **Semantic Medline**, for managing the results of PubMed searches. The portal is designed as a Java-based Web application that seamlessly integrates PubMed, **SemRep** processing of the results of the search, automatic summarization [12, 57], and, finally, visualization of the results, with links to the underlying citations and relevant additional knowledge in the UMLS Metathesaurus, the Genetics Home Reference, and Entrez Gene. We propose **Semantic Medline** as an enabling information resource for biomedical scientists, clinical decision support developers, health professionals, patients, as well as the general public.

The Advanced Library Services project has considerable potential to support health and health care. In actively constructing informatics resources for basic research and information dissemination, the **Biomedical Knowledge Repository** and associated applications can play a significant role in enabling scientific discovery, helping translate discoveries into advances in patient care, and providing the basis for individual decision making.

9 References

1. Institute of Medicine, *Crossing the quality chasm: A new health system for the 21st century*. 2001, The National Academies Press.

2. Osheroff, J., J. Teich, B. Middleton, E. Steen, B. Wright, and D. Detmer, *A roadmap for national action on clinical decision support*. 2006, American Medical Informatics Association.

3. Alfarano, C., C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M.J. Dumontier, M.R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J.P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J.J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B.F. Ouellette, and C.W. Hogue, *The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Res, 2005. **33**(Database issue): p. D418-24.

4. Thorn, C.F., T.E. Klein, and R.B. Altman, *PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base.* Methods Mol Biol, 2005. **311**: p. 179-91.

5. Micromedex. *DRUGDEX*. [cited 2006 August, 25]; Available from: http://www.micromedex.com.

6. DrugDigest. *DrugDigest*. [cited 2006 August 25]; Available from: http://www.drugdigest.org.

7. Blaschke, C., M. Andrade, C. Ouzounis, and A. Valencia, *Automatic extraction of biological information from scientific text: protein-protein interactions*, in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, T. Lenauer, et al., Editors. 1999, Morgan Kaufman Publishers, Inc: San Francisco, CA. p. 60-67.

8. Cimino, J.J. and G.O. Barnett, *Automatic knowledge acquisition from MEDLINE.* Methods Inf Med, 1993. **32**(2): p. 120-30.

9. Lussier, Y., T. Borlawski, D. Rappaport, Y. Liu, and C. Friedman, *PhenoGO: assigning phenotypic context to Gene Ontology annotations with natural language processing.* Pac Symp Bio, 2006: p. 64-75.

10. Mamlin, B.W., D.T. Heinze, and C.J. McDonald, *Automated extraction and normalization of findings from cancer-related free-text radiology reports.* AMIA Annu Symp Proc, 2003: p. 420-4.

11. Mendonca, E.A. and J.J. Cimino, *Automated knowledge extraction from MEDLINE citations.* Proc AMIA Symp, 2000: p. 575-9.

12. Fiszman, M., T. Rindflesch, and H. Kilicoglu, *Abstraction summarization for managing the biomedical research literature.* Proc HLT-NAACL Workshop on Computational Lexical Semantics, 2004: p. 76-83.

13. McKeown, K.R., S.-F. Chang, J. Cimino, S. Feiner, C. Friedman, L. Gravano, V. Hatzivassiloglou, S. Johnson, D.A. Jordan, J.L. Klavans, A. Kushniruk, V. Patel, and S. Teu-

fel, *PERSIVAL, a system for personalized search and summarization over multimedia healthcare information.* Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, 2001 p. 331-340.

14. Demner-Fushman, D., *Complex question answering based on a semantic domain model of medicine*. 2006, University of Maryland doctoral dissertation.

15. Jacquemart, P. and P. Zweigenbaum, *Towards a medical question-answering system: a feasibility study.* Stud Health Technol Inform, 2003. **95**: p. 463-8.

16. Sable, C., M. Lee, H.R. Zhu, and H. Yu, *Question analysis for biomedical question answering.* AMIA Annu Symp Proc, 2005: p. 1102.

17. Wedgwood, J., *MQAF: a medical question-answering framework.* AMIA Annu Symp Proc, 2005: p. 1150.

18. Hristovski, D., C. Friedman, and T. Rindflesch, *Exploiting semantic relations for literature-based discovery.* Proc AMIA Symp, 2006: p. (in press).

19. Swanson, D.R., *Fish oil, Raynaud's syndrome, and undiscovered public knowledge.* Perspect Biol Med, 1986. **30**(1): p. 7-18.

20. Rindflesch, T., M. Fiszman, and B. Libbus, *Semantic interpretation for the biomedical research literature*, in *Medical informatics: Advances in knowledge management and data mining in biomedicine,* H. Chen, et al., Editors. 2005, Springer-Verlag. p. 399-422.

21. Rindflesch, T.C. and M. Fiszman, *The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text.* J Biomed Inform, 2003. **36**(6): p. 462-77.

22. Masseroli, M., H. Kilicoglu, F.M. Lang, and T.C. Rindflesch, *Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease.* BMC Bioinformatics, 2006. **7**(1): p. 291.

23. Rindflesch, T.C., B. Libbus, D. Hristovski, A.R. Aronson, and H. Kilicoglu, *Semantic relations asserting the etiology of genetic diseases.* AMIA Annu Symp Proc, 2003: p. 554-8.

24. McCray, A.T., S. Srinivasan, and A.C. Browne, *Lexical methods for managing variation in biomedical terminologies.* Proc Annu Symp Comput Appl Med Care, 1994: p. 235-9.

25. Smith, L., T. Rindflesch, and W.J. Wilbur, *MedPost: a part-of-speech tagger for bioMedical text.* Bioinformatics, 2004. **20**(14): p. 2320-1.

26. Aronson, A.R., *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* Proc AMIA Symp, 2001: p. 17-21.

27. Tanabe, L. and W.J. Wilbur, *Tagging gene and protein names in biomedical text.* Bioinformatics, 2002. **18**(8): p. 1124-32.

28. Friedman, C., L. Shagina, Y. Lussier, and G. Hripcsak, *Automated encoding of clinical documents based on natural language processing.* J Am Med Inform Assoc, 2004. **11**(5): p. 392-402.

29. Leroy, G., H. Chen, and J.D. Martinez, *A shallow parser based on closed-class words to capture relations in biomedical text.* J Biomed Inform, 2003. **36**(3): p. 145-58.

30. Koike, A., Y. Niwa, and T. Takagi, *Automatic extraction of gene/protein biological functions from biomedical text.* Bioinformatics, 2005. **21**(7): p. 1227-36.

31. Hu, Z.Z., M. Narayanaswamy, K.E. Ravikumar, K. Vijay-Shanker, and C.H. Wu, *Literature mining and database annotation of protein phosphorylation using a rule-based system.* Bioinformatics, 2005. **21**(11): p. 2759-65.

32. Friedman, C., P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.* Bioinformatics, 2001. **17 Suppl 1**: p. S74-82.

33. Vizenor, L., O. Bodenreider, L. Peters, and A.T. McCray, *Enhancing biomedical ontologies through alignment of semantic relationships: Exploratory approaches.* Proc AMIA Symp, 2006: p. (in press).

34. Maglott, D., J. Ostell, K.D. Pruitt, and T. Tatusova, *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2005. **33**(Database issue): p. D54-8.

35. Blake, J.A., J.T. Eppig, C.J. Bult, J.A. Kadin, and J.E. Richardson, *The Mouse Genome Database (MGD): updates and enhancements.* Nucleic Acids Res, 2006. **34**(Database issue): p. D562-7.

36. Weng, S., Q. Dong, R. Balakrishnan, K. Christie, M. Costanzo, K. Dolinski, S.S. Dwight, S. Engel, D.G. Fisk, E. Hong, L. Issel-Tarver, A. Sethuraman, C. Theesfeld, R. Andrada, G. Binkley, C. Lane, M. Schroeder, D. Botstein, and J. Michael Cherry, *Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins.* Nucleic Acids Res, 2003. **31**(1): p. 216-8.

37. Berners-Lee, T., J. Hendler, and O. Lassila, *The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.* Scientific American, 2001. **284**(5): p. 34-+.

38. Bodenreider, O. and R. Stevens, *Bio-ontologies: current trends and future directions.* Brief Bioinform, 2006: p. bbl027.

39. World Wide Web Consortium. *Health Care and Life Sciences Interest Group*. [cited 2006 August 25]; Available from: http://www.w3.org/2001/sw/hcls/.

40. Baclawski, K. and T. Niu, *Ontologies for bioinformatics*. Computational molecular biology. 2005, Cambridge, Mass.: MIT Press.

41. Wu, C.H., R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, *The Universal Protein Resource (UniProt): an expanding universe of protein information.* Nucleic Acids Res, 2006. **34**(Database issue): p. D187-91.

42. Jain, E. *UniProt RDF*. [cited 2006 August 25]; Available from: http://expasy3.isb-sib.ch/~ejain/rdf/.

43. World Wide Web Consortium. *Health Care and Life Sciences Interest Group, BioRDF Subgroup*. [cited 2006 August 25]; Available from: http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup.

44. World Wide Web Consortium. *XSLT (eXtensible Stylesheet Language Transformation)*. [cited 2006 August 25]; Available from: http://www.w3.org/TR/xslt/.

45. World Wide Web Consortium. *GRDDL (Gleaning Resource Descriptions from Dialects of Languages)*. [cited 2006 August 25]; Available from: http://www.w3.org/TR/grddl/.

46. LSID Community. *LSID (Life Sciences Identifier)*. [cited 2006 August 25]; Available from: http://lsid.sourceforge.net/.

47. myGrid Consortium. *Taverna*. [cited 2006 August 25]; Available from: http://taverna.sourceforge.net/.

48. BioPathways Consortium. [cited 2006 August 25]; Available from: http://biopathways.org/.

49. World Wide Web Consortium. *Naming and Addressing: URIs, URLs, ...* [cited 2006 August 25]; Available from: http://www.w3.org/Addressing/.

50. Kunze, J. and R.P.C. Rodgers. *The ARK Persistent Identifier Scheme*. 2006 [cited 2006 August 25]; Available from: http://ark.cdlib.org/arkspec.pdf.

51. Sahoo, S., *Converting biological information to the W3C Resource Description Framework (RDF): Experience with Entrez Gene*. 2006, National Library of Medicine, Lister Hill National Center for Biomedical Communications.

52. World Wide Web Consortium. *XML Path Language (XPath)*. [cited 2006 August 25]; Available from: http://www.w3.org/TR/xpath.

53. Dinakarpandian, D., Y. Lee, K. Vishwanath, and R. Lingambhotla, *MachineProse: an ontological framework for scientific assertions.* J Am Med Inform Assoc, 2006. **13**(2): p. 220-32.

54. Gao, Y., J. Kinoshita, E. Wu, E. Miller, R. Lee, A. Seaborne, S. Cayzer, and T. Clark, *SWAN: A distributed knowledge infrastructure for Alzheimer disease research.* Journal of Web Semantics, 2006. **4**(3).

55. OpenRDF. *Sesame*. [cited 2006 August 25]; Available from: http://www.openrdf.org/.

56. World Wide Web Consortium. *SPARQL*. [cited 2006 August 25]; Available from: http://www.w3.org/TR/rdf-sparql-query/.

57. Fiszman, M., T. Rindflesch, and H. Kilicoglu, *Summarizing drug information in Medline citations.* Proc AMIA Symp, 2006: p. (in press).

58. Card, S.K., J.D. Mackinlay, and B. Shneiderman, *Readings in information visualization : using vision to think*. The Morgan Kaufmann series in interactive technologies. 1999, San Francisco, Calif.: Morgan Kaufmann Publishers. xvii, 686 p.

59. Plake, C., T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser, *ALIBABA: PubMed as a graph.* Bioinformatics, 2006.

60. Fuller, S.S., D. Revere, P.F. Bugni, and G.M. Martin, *A knowledgebase system to enhance scientific discovery: Telemakus.* Biomed Digit Libr, 2004. **1**(1): p. 2.

61. Jenssen, T.K., A. Laegreid, J. Komorowski, and E. Hovig, *A literature network of human genes for high-throughput analysis of gene expression.* Nat Genet, 2001. **28**(1): p. 21-8.

62. van der Eijk, C., E. van Mulligen, J. Kors, and B. Mons, *Constructing an associative concept space for literature-based discovery.* JASIST, 2004. **55**(5): p. 436-44.

63. Feldman, R., Y. Regev, E. Hurvitz, and M. Finkelstein-Landau, *Mining the biomedical literature using semantic analysis and natural language processing techniques.* Biosilico, 2003. **1**(2): p. 69-80.

64. Tao, Y., C. Friedman, and Y.A. Lussier, *Visualizing information across multidimensional post-genomic structured and textual databases.* Bioinformatics, 2005. **21**(8): p. 1659-67.

65. Marshall, B.J. and H.M. Windsor, *The relation of Helicobacter pylori to gastric adenocarcinoma and lymphoma: pathophysiology, epidemiology, screening, clinical presentation, treatment, and prevention.* Med Clin North Am, 2005. **89**(2): p. 313-44, viii.

66. Tsarkov, D. and I. Horrocks, *Efficient reasoning with range and domain constraints.* Proc. DL 2004, 2004: p. 41–50.

67. Bodenreider, O., *Circular hierarchical relationships in the UMLS: Etiology, diagnosis, treatment, complications and prevention.* Proc AMIA Symp, 2001: p. 57-61.

68. Cimino, J.J., *Auditing the Unified Medical Language System with semantic methods.* J Am Med Inform Assoc, 1998. **5**(1): p. 41-51.

69. McCray, A.T. and O. Bodenreider, *A conceptual framework for the biomedical domain*, in *The semantics of relationships: an interdisciplinary perspective*, R. Green, C.A. Bean, and S.H. Myaeng, Editors. 2002, Kluwer Academic Publishers: Boston. p. 181-198.

70. McGuire, W., ed. *British medical journal: Clinical evidence concise*. 2004, BMJ Publishing Group: London.

71. First DataBank. *National Drug Data File.* [cited 2006 August 25]; Available from: http://www.firstdatabank.com/.

72. Mani, I., *Automatic summarization*. 2001, Amsterdam ; Philadelphia: J. Benjamins Pub. Co.

73. McKeown, K., R.J. Passonneau, D.K. Elson, A. Nenkova, and J. Hirschberg, *Do summaries help?* Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005 p. 210-217.

10 Appendix

10.1 Appendix A: Example of XML representation

The Entrez Gene XML representation of the proteins coded by Gene with geneid 351 (representative fragment of XML with extra element tags to be valid XML)

```
<?xml version="1.0"?>
<!DOCTYPE Entrezgene-Set PUBLIC "-//NLM//DTD NCBI-Entrezgene, 21st January 2005//EN" "NCBI_Entrezgene.dtd">
<Entrezgene-Set>
<Entrezgene>
<Entrezgene_track-info>
    <Gene-track>
      <Gene-track_geneid>351</Gene-track_geneid>
    </Gene-track>
</Entrezgene_track-info>
<Entrezgene_prot>
    <Prot-ref>
      <Prot-ref_name>
        <Prot-ref_name_E>amyloid beta A4 protein</Prot-ref_name_E>
        <Prot-ref_name_E>protease nexin-II</Prot-ref_name_E>
        <Prot-ref_name_E>A4 amyloid protein</Prot-ref_name_E>
        <Prot-ref_name_E>amyloid-beta protein</Prot-ref_name_E>
        <Prot-ref_name_E>beta-amyloid peptide</Prot-ref_name_E>
        <Prot-ref_name_E>cerebral vascular amyloid peptide</Prot-ref_name_E>
        <Prot-ref_name_E>amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)</Prot-ref_name_E>
      </Prot-ref_name>
    </Prot-ref>
  </Entrezgene_prot>
  </Entrezgene>
</Entrezgene-Set>
```

10.2 Appendix B. Example of RDF representation

The Entrez Gene RDF representation of the proteins coded by Gene with *geneid* 351 (representative fragment of RDF with extra element tags to be valid XML).

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:eg="http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/">
<rdf:Description rdf:about="eg:Gene-track_geneid/351">
<eg:has_entrezgene_protein rdf:parseType="Resource">
    <eg:has_protein_reference rdf:parseType="Resource">
      <eg:has_protein_reference_name rdf:parseType="Resource">
        <eg:has_protein_reference_name_E>amyloid beta A4 protein</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>protease nexin-II</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>A4 amyloid protein</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>amyloid-beta protein</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>beta-amyloid peptide</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>cerebral vascular amyloid peptide</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer
disease)</eg:has_protein_reference_name_E>
      </eg:has_protein_reference_name>
    </eg:has_protein_reference>
  </eg:has_entrezgene_protein>
  </rdf:Description>
</rdf:RDF>
```