

Integration of Relational and Textual Biomedical Sources

A Pilot Experiment Using a Semi-automated Method for Logical Schema Acquisition

M. García-Remesal¹; V. Maojo¹; H. Billhardt²; J. Crespo¹

¹Biomedical Informatics Group, Dep. Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain;

²Artificial Intelligence Group, Universidad Rey Juan Carlos, Madrid, Spain

Keywords

Database integration, genetic diseases, text mining, natural language processing

Summary

Objectives: Bringing together structured and text-based sources is an exciting challenge for biomedical informaticians, since most relevant biomedical sources belong to one of these categories. In this paper we evaluate the feasibility of integrating relational and text-based biomedical sources using: i) an original logical schema acquisition method for textual databases developed by the authors, and ii) OntoFusion, a system originally designed by the authors for the integration of relational sources.

Methods: We conducted an integration experiment involving a test set of seven differently structured sources covering the domain of genetic diseases. We used our logical schema acquisition method to generate schemas for all textual sources. The sources were integrated using the methods and tools provided

by OntoFusion. The integration was validated using a test set of 500 queries.

Results: A panel of experts answered a questionnaire to evaluate i) the quality of the extracted schemas, ii) the query processing performance of the integrated set of sources, and iii) the relevance of the retrieved results. The results of the survey show that our method extracts coherent and representative logical schemas. Experts' feedback on the performance of the integrated system and the relevance of the retrieved results was also positive. Regarding the validation of the integration, the system successfully provided correct results for all queries in the test set.

Conclusions: The results of the experiment suggest that text-based sources including a logical schema can be regarded as equivalent to structured databases. Using our method, previous research and existing tools designed for the integration of structured databases can be reused – possibly subject to minor modifications – to integrate differently structured sources.

In this scenario, we earlier completed the INFOGENMED project [7], targeted to the integration of clinical and genomic structured databases. In this context, we developed the OntoFusion system [8], a suite of tools to integrate and query clinical and genomic structured data repositories following a mediator/wrapper-based approach [9]. Relational and object-oriented databases are the most representative examples of structured sources.

Over the last few years, biomedical researchers have been showing a growing interest in another type of resources, namely text-based sources. The latter can be regarded as large text-based collections that do not have a logical schema describing their contents. Examples of text-based databases range from collections of OCR'ed (Optical Character Recognition) clinical records to large online biomedical databases such as MEDLINE or OMIM.

Unfortunately, the methods and tools provided by OntoFusion cannot be reused to bridge together structured and text-based sources. Furthermore, all major state-of-the-art approaches to structured database integration require the individual sources to be equipped with a logical schema.

To address this issue, we propose the automated extraction of logical schemas from text-based sources using an original logical schema acquisition method developed by the authors and reported elsewhere [10]. Once a logical model has been obtained for a given textual resource, the latter can be regarded as equivalent to a structured database. Hence, readily available methods and tools designed to integrate structured sources can be reused to bridge together sets of data resources involving both structured and text-based sources.

Methods Inf Med 4/2010

Correspondence to:

Miguel García-Remesal, PhD
Dep. Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid
Campus de Montegancedo S/N
28660 Boadilla del Monte
Madrid
Spain
E-mail: mgremesal@fi.upm.es

Methods Inf Med 2010; 49: 337–348

doi: 10.3414/ME0614

received: November 11, 2008

accepted: August 11, 2009

prepublished: November 20, 2009

1. Introduction

The results of omics' research have generated a wealth of data scattered across different databases distributed all over the world. This situation is generating new and excit-

ing challenges for biomedical informaticians [1–4]. This includes the development of new methods and tools to collect, integrate and retrieve all the available online information focused on biomedical topics [5, 6].

We evaluated the feasibility of the proposed approach, by carrying out an integration experiment using a test set including seven differently structured sources covering the domain of genetic diseases. A panel of experts was convened to provide feedback and quality assessment on the generated logical schemas for the text-based sources. The correctness of the integration was also validated using a test set composed of 500 queries.

2. Background

Over the last years, two major methodologies have been extensively used to address the integration of heterogeneous sources [11]: i) data translation and ii) query translation. The first approach, which aims to integrate structured sources, relies on the creation of a centralized repository whose logical structure models the domain covered by the set of target sources. Data belonging to the different databases are converted into a common format and stored in the centralized repository. Data warehousing (DW) [12] is the most representative technique following the data translation approach. A concrete implementation of DW in the biomedical domain is reported in [13].

Conversely, the query translation methodology relies in the use of mediators [9] and wrappers to query an integrated view of the sources. Conversely to the data translation approach – in which actual data is replicated in a centralized repository – data are stored in the original sources. Queries launched against the integrated schema – i.e. involving conceptual objects belonging to the conceptual view – are decomposed into sub-queries and forwarded to the physical sources by the mediator. These sub-queries are then translated by the wrappers to i) use objects belonging to the logical schemas of the underlying sources, and ii) match the corresponding query language for each source. Results are translated into conceptual entities by the wrappers and then sent to the mediator that unifies and presents the results to the user. This approach has been extensively used in recent database integration systems [8, 14–18].

Early mediation-based systems such as TSIMMIS [14] do not use integrative domain models to facilitate schema reconciliation. Instead, they focus on the generation of complex wrappers to query individual sources based on their contents. Each source is assigned a custom-made wrapper that is regarded as an expert in the contents of its respective source. The main drawback

of these systems is that users find it hard to formulate queries, since they do not have an available model describing the contents of the underlying sources. To use the system, users are required to have a deep knowledge on the domain of interest.

Most recent heterogeneous source integration systems have adopted ontology-based approaches. As reported in the literature, together with mediators and wrappers, ontologies have proven to be a useful resource for information integration tasks [8, 15, 18–20]. Ontologies help users to understand the domain covered by the sources, facilitating the query formulation process.

In the biomedical domain, recent work on ontology-based database integration includes systems such as TAMBIS [18], BACIIS [19], and OntoFusion [8]. However, most of these systems only provide support for one type of sources: either structured or text-based. We recently updated OntoFusion to provide support for public online gene, protein and disease resources [21]. The approach we took was based on automatically creating a relational database for each text-based resource by extracting the relevant data from the actual web pages.

We believe that methods and tools designed for structured database integration could be reused to integrate structured and text-based sources if the latter were equipped with a logical schema. If such schemas could be extracted automatically, then the text-based sources could be regarded as being equivalent to structured sources. Thus, readily available methods and tools for structured database integration could be reused to integrate differently structured sources.

3. Methods

3.1 A Four-phased Method for Logical Schema Acquisition

As shown in ►Figure 1, the schema acquisition method is composed of four activities: i) vocabulary extraction, ii) taxonomical organization of the terminology, iii) “ad-hoc” relationships discovery, and iv) optional schema refinement. A detailed description of each phase follows.

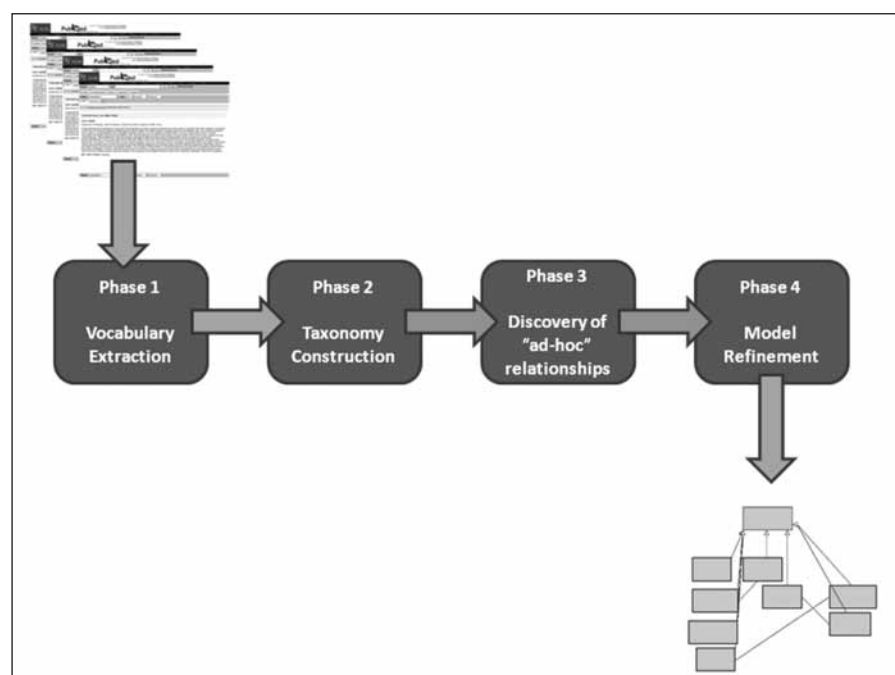


Fig. 1 Overview of the logical schema extraction process

3.1.1 Phase 1: Vocabulary Extraction

The objective of phase 1 is to discover all relevant concepts contained in the documents of the collection.

In this activity, the documents are processed to extract all concepts or classes that may be relevant to the domain of interest. To identify concepts in the text, we search for noun phrases (NPs). The latter can be defined as sequences of one or more words that together can serve as the subject or the object of a verb.

Many different methods can be applied to identify NPs in free text. They include statistical and natural language processing (NLP) techniques, and also combinations of both approaches. We decided to use a combination of different NLP techniques, since these approaches normally perform well enough for NP extraction. However, any other available technique or algorithm can be used – e.g. the National Library of Medicine's MetaMap [22] algorithm.

To extract the relevant vocabulary contained in the textual collection, we use an original algorithm developed by the authors. The algorithm relies on several existing components, such as lexical analyzers or part-of-speech (POS) taggers, and on a vocabulary server developed by the authors. The description of the algorithm follows.

First, we divided each document into phrases using a phrase generator. The latter can recognize the beginning and ending of sentences, and splits the document into a set of phrases. Each sentence is then tokenized using a simple scanner (lexical analyzer), to produce a set of tokens (words). Tokens are then labeled with their respective POS tags applying a widely used probabilistic and language-independent POS tagger called QTAG [23, 24].

Once the phrase has been successfully converted into a succession of POS tags, then we moved on to detect the NPs it contains. To carry out this task, we perform a syntactic analysis based on the use of transition networks [25] (TNs). We used three different TNs to detect three different types of NPs: simple, conjunctive/disjunctive, and adverbial NPs. Simple NPs are composed of zero or more adjectives followed by at least one common or proper

noun. Some examples of simple NPs are “protein”, “organic compound” or “guided tissue regeneration”. By contrast, conjunctive/disjunctive NPs are built upon conjunctions or disjunctions of simple NPs. Some examples are “Low T3 and T4 Syndrome” or “B and T Cell Acute Lymphoblastic Leukemia”. Finally, adverbial NPs are composed of an adverbial form followed by a simple or conjunctive NP. An example of adverbial NP is “badly injured arm”.

The execution of the different TNs produces a set of candidate NPs that needs to be evaluated to discard irrelevant concepts. To assess the relevance of a given NP, we search for it in a biomedical terminology server powered by OntoFusion. This server integrates terminology from several biomedical vocabularies, including the Unified Medical Language System [26, 27] (UMLS), the Gene Ontology [28, 29] (GO) – currently subsumed in the UMLS –, the Human Gene Nomenclature [30, 31] (HGN), and the Foundational Model of Anatomy [32, 33] (FMA). Regarding the vocabulary server, further details on its structure and functioning can be found elsewhere [8].

Once the search has been conducted, if the target NP is found, then it is marked as relevant. Next, it is automatically linked to a list of semantically similar terms (i.e. synonyms) provided by the vocabulary server. After that, the concept is assigned its preferred string according to the database contents. If the concept cannot be found in the vocabulary server then it is marked as discarded, and it is not considered for the following activities. However, if necessary, it can be re-included in the logical model by experts in the refinement phase.

After the pruning, we calculate the CF-IDF (Concept Frequency-Inverse Document Frequency) score proposed by Hersh et al. [34] for each of the remaining NPs. This score is a popular measure of the importance of a concept. It evaluates the trade-off between the importance of a concept in the context of a single document and in the context of the whole collection. This information can be used to prune the set of remaining concepts. By setting a threshold, all concepts whose CF-IDF value is smaller than the previously selected cut-

off value are discarded. This parameter has to be tuned to match the desired size of the logical model – in terms of the number of concepts.

The result of this phase is a set of concepts containing the most relevant terms found in the textual collection.

3.1.2 Phase 2: Taxonomy Construction

The second activity of our method aims to automatically discover hyponymy relationships between pairs of concepts contained in the extracted vocabulary. Hyponymy relationships are those that relate a generic concept – or class of objects – to its more specific terms – or subclasses. To carry out this task, we followed the approach described below.

First of all, we generate all different pairs of concepts belonging to the extracted vocabulary. Then, we search the vocabulary server provided by OntoFusion to determine which of these pairs of terms are actually linked by an “*is a*” or similar hierarchical relationship. All pairs of concepts, according to the vocabulary server, connected by a “*is a*” relationship are marked as components of the taxonomy of concepts of the target logical schema.

The vocabulary server provided by OntoFusion determines whether a given pair of concepts are connected by an “*is a*” link according to the most relevant biomedical terminologies – e.g. UMLS, GO, and the FMA. Even so, we think that the discarded pairs of concepts could still include interesting hierarchical relationships not recognized by the vocabulary server. To address this issue we followed a pattern-matching method based on Hearst's approach [35].

We manually built a knowledge base containing around 90 hyponymy patterns. We included patterns manually extracted by the authors from i) elementary medical and biological textbooks and web pages and ii) several hundred abstracts of biomedical research papers selected from PubMed. We also included into the knowledge base the six hyponymy patterns proposed by Hearst [35].

The pattern-matcher resorts to a rule base inference engine built using JESS [36], the Java-based rule engine. It uses an en-

hanced version of the RETE [37] algorithm to process the rules. Hyponymy patterns were converted into rules to find instances of patterns in the texts using JESS. ▶ Table 1 shows some sample rules together with examples of hyponymy relationships that can be extracted using such rules.

The knowledge base also includes some “rule of thumb” patterns that are applicable in any domain. This accounts for rules such as “If an NP is composed by N words, then the NP composed by the last N-1 words of the former NP has a broader meaning than the original”. This “rule of thumb” provides hyponymy relationships such as “Neutron Tomography is a Tomography”.

The result of this activity is a hierarchy of concepts built upon the discovered hyponymy links. All orphan nodes – i.e. those not connected by hyponymy relationships – are automatically connected to an artificial root node. Terms located at the highest level in the taxonomy are connected to the root node as well.

After building the taxonomy, we can move on the next phase.

3.1.3 Phase 3: Discovery of “Ad-hoc” Relationships

Phase 3 aims to discover non-standard or *ad-hoc* relationships between con-

cepts in the previously created hierarchy.

Ad-hoc relationships represent non-hierarchical links between objects, such as for instance “Virus causes Disease” or “Disease is caused by Virus”. In this example, both sentences reflect the same conceptual relationship. However, the verbs “causes” and “is caused by” denote the roles played by the first concept with respect to the second concept.

There are two different types of *ad-hoc* relationships: class-class and class-attribute relationships. Class-class relationships denote a connection between two conceptual entities, such as for instance “Codon encodes Aminoacid”. Conversely, class-attribute relationships represent links between an entity and one of its attributes. For example, as codons are composed of a sequence of three sorted bases, the *Sequence* can be regarded as an attribute of the *Codon*.

The method we propose to discover *ad-hoc* relationships does not provide role names for the discovered relations, nor does it distinguish between class-class and class-attribute relationship types. However, it provides connections between concepts in the hierarchy together with a measure of the reliability of the extracted relationship. Role names can be assigned by experts in the refinement phase if required.

To discover *ad-hoc* relationships, we propose an approach based on collocations [38]. Collocations can be defined as short-distance occurrences of two or more words in a given context. This context may range from a whole textual document to a single sentence.

We modified the original collocations-based approach described in [38] to use concepts rather than words. Thus, we did not use the original text documents from the collection to compute the collocations. Instead, we replaced each document *D* with a list *L_D* of relevant concepts. Relevant concepts are those that belong to the hierarchy of concepts created in the previous phase. Concepts are sorted by order of appearance in the original document.

To determine whether there is a relationship between a pair of concepts applying the collocations approach, we tested the following null hypothesis: “Given two concepts, the occurrence of one of them is independent of the occurrence of the other in the same text”. To conduct this hypothesis testing we use the t-score [38] statistic, based on Student’s t distribution [39]. This measure compares the observed and expected frequencies of co-occurrences between both concepts.

To calculate observed co-occurrence frequencies, we use small contexts called

Rule	Sample sentence	Extracted hyponyms
$NP1 \text{ is } \{a \mid an\} \text{ } NP0 \rightarrow \text{Hyponym}(NP1, NP0)$	« Achondroplasia is a genetic disorder that... »	Hyponym(achondroplasia, genetic disorder)
$NP1 \text{ is a } \{type \mid kind\} \text{ of } NP0 \rightarrow \text{Hyponym}(NP1, NP0)$	« ...A double-blinded trial is a type of clinical trial in which... »	Hyponym(double-blinded trial, clinical trial)
$NP1 \{, \mid NPi\}^* \{and \mid or\} \text{ other } NP0 \rightarrow \text{Hyponym}(\text{SingularOf}(NPk), \text{SingularOf}(NP0)) \forall k = 1...i$	« ... intended for use primarily by physicians, biologists and other health professionals concerned with genetic disorders... »	Hyponym(physician, health professional), Hyponym(biologist, health professional)
$NP0 \text{ such as } \{for \text{ instance } \mid for \text{ example}\} NP1 \{, \mid and \mid or\} NPi\}^* \rightarrow \text{Hyponym}(\text{SingularOf}(NPk), \text{SingularOf}(NP0)) \forall k = 1...i$	« In the case of neuroectodermal tumors, such as for instance neuroblastoma, primitive peripheral neuroectodermic tumor, Ewing’s sarcoma... »	Hyponym(neuroblastoma, neuroectodermal tumor), Hyponym(primitive peripheral neuroectodermic tumor, neuroectodermal tumor), Hyponym(Ewing’s sarcoma, neuroectodermal tumor)
$NP0 \{, \} \text{ especially } \{NPi\}^* \{or \mid and\} NP1+1 \rightarrow \text{Hyponym}(\text{SingularOf}(NPk), \text{SingularOf}(NP0)) \forall k = 1...i+1$	« ...in the presence of organic substances, especially proteins.»	Hyponym(protein, organic substance)
$NP1 \text{ is a common } \{type \mid kind\} \text{ of } NP0 \rightarrow \text{Hyponym}(NP1, NP0)$	« Alport syndrome is a common type of hereditary glomerulopathy »	Hyponym(Alport syndrome, hereditary glomerulopathy)

Table 1
Sample rules to match hyponymy patterns

concordance lines. The latter can be defined as ordered sequences of concepts of size $2s + 1$ centered on a given concept. This is called node concept. The parameter s is a span factor that determines the size of the context. Typical values for s range from 5 to 15.

To determine whether an *ad-hoc* relationship holds between two given concepts c_1 and c_2 we proceed as follows. First, we calculated all concordance lines centered on concept c_1 across all documents in the collection. The result was a set of concordance lines whose node concept is c_1 . The observed co-occurrence frequency for concepts c_1 and c_2 was then computed as the number of occurrences of concept c_2 in the set of concordance lines associated with c_1 . Expected concept co-occurrence frequencies were computed analogously from a large corpus extracted from the PubMed database.

Once observed and expected concept co-occurrences have been estimated, it is possible to compute the value of the t-score statistic. Values greater than 2 indicate that we must reject the null hypothesis, thus entailing a relationship between the concepts.

This activity produces a symmetric square matrix R of size n , n being the number of different concepts in the taxonomy of concepts. Each element $c_{i,j}$ in the matrix indicates the strength of the relationship between concepts c_i and c_j . Pairs of concepts whose t-score statistic is greater than 2 are then automatically linked with a relationship. However, note that relationships are not assigned role names. The latter can be extracted by replacing the collocation-based detector with an extractor of semantic predications such as SemRep [40] or BioMedLee [41].

3.1.4 Phase 4: Model Refinement

Phase 4 is targeted to the refinement of the extracted logical model. It is a manual activity that must be conducted by domain experts assisted by a knowledge engineer. Activities to be performed during this phase include i) the addition or removal of concepts and relationships, and ii) the enhancement of the extracted model using other available biomedical terminologies. Data and statistics

Table 2 Summary of resources selected by the panel of experts

Resource	Type	Description
Clinical trials	Text-based	Information on a clinical trial's purpose, participants, locations, phone numbers, etc. Available at http://clinicaltrials.gov
EDDNAL	Text-based	Information on molecular diagnostic services for heritable syndromes and disorders provided by European laboratories. Available at http://www.eddnal.com
OMIM	Text-based	A database of human genes and genetic disorders. Available at http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
Orphanet	Text-based	A database that provides information on rare disease and orphan drugs. Available at http://www.orpha.net/
PharmGKB	Text-based	Provides information on the relationships among drugs, diseases and genes, including their variations and gene products. Available at http://www.pharmgkb.org
PubMed	Text-based	A database of abstracts of biomedical scientific articles. Provides links to full text articles and other related resources. Available at http://www.ncbi.nlm.nih.gov/PubMed
RDClinic	Structured (relational)	Relational database developed in the context of the INFO-GENMED Project. Provides clinical data related to 185 patients diagnosed with rare genetic diseases. The database is not publicly available.

collected during the previous phases – e.g. CF-IDF scores and the R matrix – together with other available resources such as the vocabulary server, can be used by curators during this activity. It is also possible for experts to manually assign role names to the extracted relationships if required. All these modifications can be performed by using an integrated software tool.

The result of this process is not a logical schema in the traditional sense, since it does not include elements such as tables, attributes and relationships. Instead, it includes concepts and relationships describing the information contained in the textual source.

3.2 Description of the Integration Experiment

To prove the feasibility of the proposed approach, we carried out an integration experiment involving a set of real world databases of different types covering the domain of genetic diseases.

First, a panel of experts – composed by three senior biomedical researchers and a computer scientist – identified a set of seven databases containing relevant data for the domain of genetic diseases. Experts were selected on the basis of their expertise and experience in i) rare genetic disorders, ii) conducting frequent searches on the selected public online databases, and iii) the use of the mapping tool provided by OntoFusion.

Six of the selected resources were public online databases, while the remaining source was a relational database containing clinical data related to patients diagnosed with rare genetic diseases. This database was developed by medical experts from the Institute of Health Carlos III, in Spain. ► Table 2 summarizes the main features of the selected databases.

The panel of experts selected a set of 200 genetic diseases based on the number of online sources providing information for each disease. Only genetic disorders covered by at least three of the selected online sources were considered in the study.

Table 3 Main features of the query test set

Query type	Sample query	# of queries
Simple queries	Q = {<Pathology, Name, contains, "McArdle disease">}	50
Average queries	Q = {<Pathology, Name, contains, "Fragile X syndrome">, <Pathology, Related_to, References>}	400
Complex queries	Q = {<Patient, ID, is, "09234593">, <Patient, related_to, Genetic Test>}	50

Next, we launched 200 queries (one per disease) thus retrieving 500 documents for each online source. If more than 500 documents were retrieved for a given database, we kept at least two for each disease when possible. The rest, up to 500 documents, was randomly selected among the remaining documents. We believe that a sample of 500 documents per source is representative enough to extract the logical structure of any domain-specific text-based source. For instance, this setting is suitable for extracting from the OMIM database concepts such as "Clinical Features" or "Inheritance",

since these generic objects occur in nearly all documents in the collection. Conversely, extracting instance-specific objects such as "Axenfeld Anomaly" or "Machado-Joseph Disease" would involve a larger number of documents – most probably the full distribution. As we are interested in obtaining logical schemas mostly including generic objects – i.e. classes rather than instances – we decided to set a small number of documents per collection compared to the actual size of the online sources.

The generated sets of documents were used to automatically build the logical

schemas of the text-based sources using the original schema acquisition method developed by the authors [10].

After the extraction of the logical schemas for all text-based sources, we used the tools provided by OntoFusion to create a unified domain model (UDM) covering the domain described by the integrated set of sources. The creation of the UDM using the methods and tools provided by OntoFusion is supported by two basic operations: mapping and unification. The mapping process [8] is targeted to translate the logical schema of the source into a mapping domain model (MDM). The latter describes the domain covered by its associated source.

On the other hand, the unification phase is aimed to merge the conceptual schemas of the individual sources to build a unified domain model (UDM). This task is carried out using an automated unification algorithm. Further details regarding the mapping and unification processes are reported elsewhere [8].

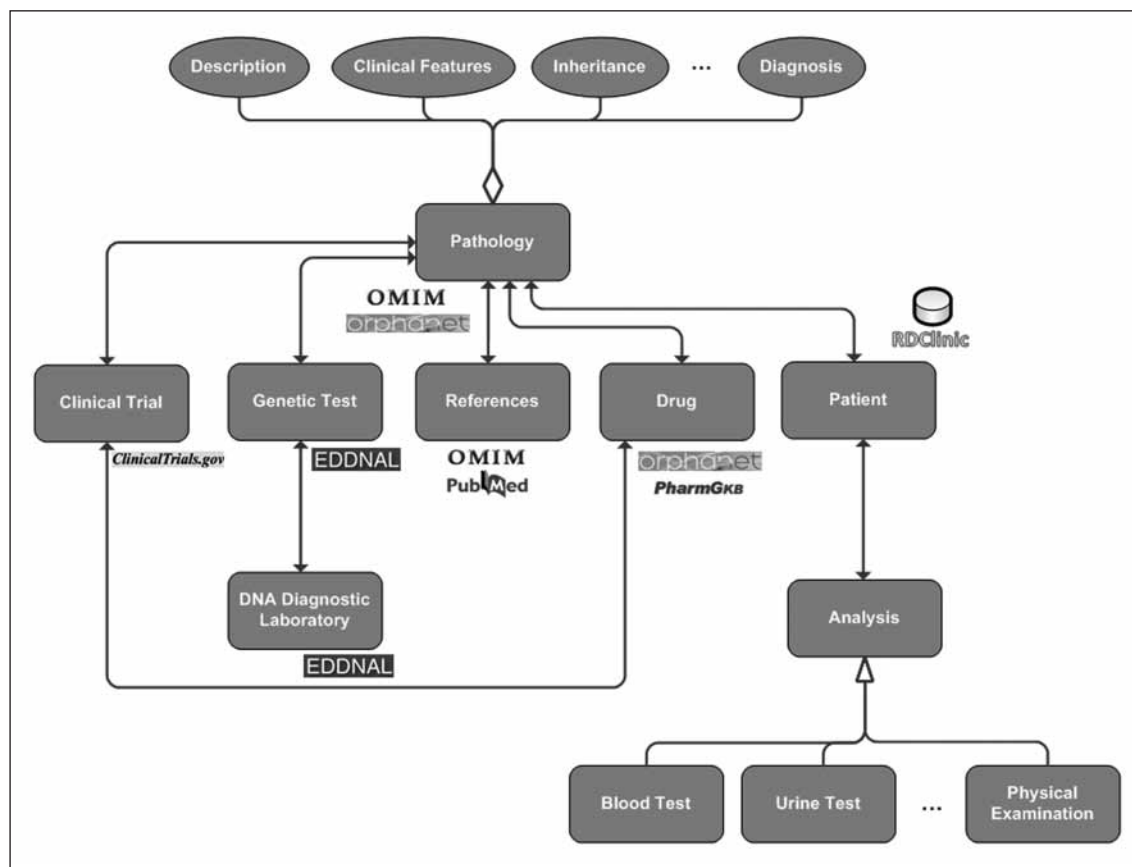


Fig. 2 Extract of the UDM that integrates all selected sources

Table 4

Summary of results of the logical schema extraction process and the mapping task. The greyed cells show the selected threshold settings for each source and the number of extracted concepts. Additional threshold values for each collection (no thresholding and $\tau = 0.3$) are presented for reference.

Source	Concepts			Hierarchical relationships		AD-HOC relationships	
	CF-IDF threshold	Extracted	Mapped	Extracted	Mapped	Extracted	Mapped
Clinical trials	None	16571	158 89.77%	125	114 91.2%	473	100 21.14%
	$\tau = 0.3$	5447					
	$\tau = 0.71$	176					
EDDNAL	None	15291	154 92.77%	83	76 91.56%	215	100 46.51%
	$\tau = 0.3$	6351					
	$\tau = 0.67$	166					
OMIM	None	32760	208 94.11%	136	127 93.38%	846	100 11.82%
	$\tau = 0.3$	9126					
	$\tau = 0.79$	221					
Orphanet	None	23675	210 98.59%	124	117 94.35%	924	100 10.82%
	$\tau = 0.3$	9547					
	$\tau = 0.71$	213					
PharmGKB	None	9420	144 91.13%	60	51 85%	186	100 53.76%
	$\tau = 0.3$	5664					
	$\tau = 0.57$	158					
Pubmed	None	21467	191 87.21%	142	140 98.59%	1015	100 9.85%
	$\tau = 0.3$	12430					
	$\tau = 0.74$	219					
RDclinic	N/A	13 + 82	95 100%	4	4 100%	27	27 100%

To validate the correctness of the integration – in terms of software validation – we used a test set composed of 500 queries. As shown in ►Table 3, the latter contains three different types of queries: i) simple queries, ii) average queries, and iii) complex queries. Simple queries are composed of a single concept and one or more attribute filters. The sample query shown in ►Table 3 can be read as “retrieve all the instances of the class Pathology whose name contains the string ‘McArdle disease’”. Our test set contains 50 queries belonging to this category. On the other hand, intermediate queries include a target concept, one or more directly related concepts and attribute filters. Directly related concepts

are those connected to the target class by a relationship belonging to the UDM. The sample average query shown in ►Table 3 can be read as “retrieve all instances of the class Pathology whose name contains the string ‘Fragile Syndrome X’, and all their related instances of class References”. Our test collection includes 400 queries of this type. Regarding complex queries, they involve a target concept, any number or directly related concepts, at least one indirectly related concept and attribute filters. Indirectly related concepts are those linked to a helper concept not considered in the query that is connected to the target concept either directly or indirectly. The sample complex query shown in ►Table 3 could be read as

“retrieve all instances of class Patient whose id is the string ‘09234593’ and all their related instances of class Genetic Test”. As shown in ►Figure 2, the concept Patient is indirectly related to the concept Genetic Test through the concept Pathology. These queries can be executed from the OntoFusion user interface in N steps, being N the number of traversed relationships from the target (origin) concept to the related (destination) concept. The test set contains 50 complex queries.

We developed a Python script to fully automate the validation activity. For each query, the script first determines which databases contain relevant information – i.e. instances of the target concept. Next, the script retrieves all instances of the target

Table 5 Summary of time needed for building the logical schemas and their associated MDMs

Source	Vocabulary extraction	Taxonomy generation	Ad-hoc relationships discovery	Refinements and mapping	Total time
Clinical trials	<1 min	<5 min	<6 min	<15 min	<27 min
EDDNAL			<7 min	<15 min	<28 min
OMIM			<9 min	<20 min	<35 min
Orphanet			<8 min	<20 min	<34 min
PharmGKB			<4 min	<10 min	<20 min
Pubmed			<8 min	<20 min	<24 min
RDclinic	N/A	N/A	N/A	<10 min	<10 min

concept meeting the user query from the previously selected sources. Queries are executed using the native query services provided by the sources. This emulates the functioning of OntoFusion when processing simple queries. When testing average and complex queries, once the instances of the target concept have been retrieved, the script randomly selects one of the retrieved instances (I) and one of the related concepts included in the query (C). Next, the script retrieves all the instances of C related to I . The script records the unique identifier of C together with the chosen related concept I to fully reproduce the same execution

path when querying the UDM. After launching the query to the UDM, the script compares the results obtained with both methods. If both result sets contain the same instances, then we consider that the query has been successfully processed by the UDM.

4. Results

4.1 The Extracted Logical Schemas and Their Associated MDMs

► Table 4 presents a numerical description of the generated logical schemas and

their associated MDMs. The table shows the total number of extracted (and mapped) concepts, hierarchical links and *ad-hoc* relationships for each text-based source. As shown in the table, we evaluated several settings for the CF-IDF threshold (τ) during the vocabulary extraction phase. We determined the thresholds for the collections by trial and error using the software tool developed by the authors. As shown in ► Table 4, we tried different settings for this value ranging from no thresholding to more constraining values until we obtained the desired number of concepts – i.e. less than 250 concepts.

Since the number of extracted *ad-hoc* relationships for each text-based source was too large to be manually refined, we decided to include only a reduced number of them into the MDMs. For each source, we presented the experts with the extracted relationships sorted in descending order of the t -score. Next, they included into the MDM the 100 relationships that they judged to be more relevant.

► Table 5 shows the time required for extracting the logical schemas and creating the MDMs for each source.

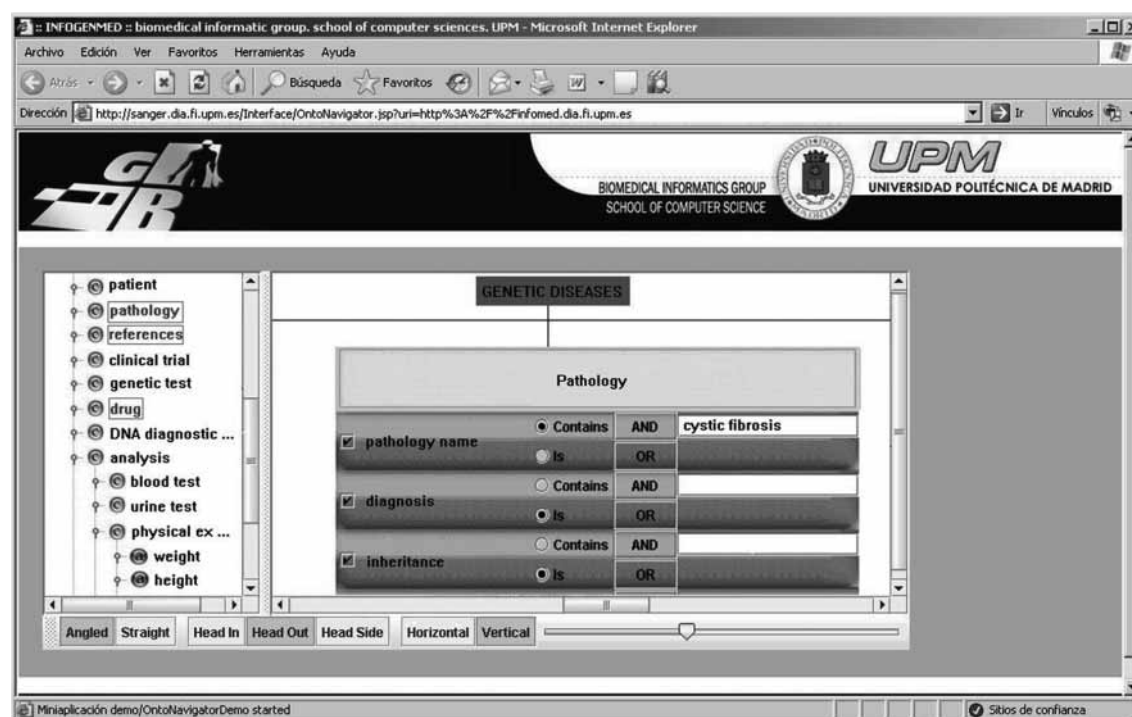


Fig. 3 Sample query launched to the set of integrated sources

The screenshot displays a web application interface. On the left, a 'pathology' list contains the following items: RDClinic_37, RDClinic_111, RDClinic_245, RDClinic_320, ORPHANET_15, ORPHANET_23, ORPHANET_127, ORPHANET_276, OMIM_602421, OMIM_219700, OMIM_178500, OMIM_174000, OMIM_123885, OMIM_603855, OMIM_219721, OMIM_123886, OMIM_263200, OMIM_208540, OMIM_603643, OMIM_277180, OMIM_190180, OMIM_173900, OMIM_208500, OMIM_256100, and OMIM_243910. A large grey arrow points from this list to the main content area. The main content area shows a 'Concept' window for 'Cystic fibrosis' with the following details: pathology name: Cystic fibrosis; diagnosis: no available information; inheritance: no available information; description: Cystic fibrosis is a genetic disorder among ... Below this are buttons for 'Related to references' and 'Related to drug'. Two browser windows are overlaid on the bottom right. The top one is the ORPHANET website with a search for 'DISEASE: Cystic fibrosis'. The bottom one is the OMIM database entry for '602421 CYSTIC FIBROSIS TRANSMEMBRANE CONDUCTANCE REGUL...', showing alternative titles like 'ATP-BINDING CASSETTE, SUBFAMILY C, MEMBER 7; ABCCT7' and a description of the CFTR gene.

Fig. 4
Results of the
sample query

4.2 The Unified Domain Model

The integrated UDM is composed of 261 concepts, 106 hierarchical relationships and 387 *ad-hoc* relationships. These figures compared to those shown in ▶ Table 4 suggest that there is a considerable overlap among the logical schemas of the sources.

▶ Figure 2 shows an extract of the UDM that integrates all the sources selected by the panel of experts. Note that only the most relevant concepts and relationships have been included in the diagram. As shown in ▶ Figure 2, three types of semantic links have been considered to create the UDM: class-attribute (marked with a diamond), hyponymy (represented by a

triangle), and *ad-hoc* relationships. Class-attribute relationships connect a given concept to other concepts – represented as ovals – that were defined as its attributes during the mapping process. For instance, the concepts *Description*, *Clinical Features*, *Inheritance* and *Diagnosis* are connected to the *Pathology* concept through class-attribute relationships.

4.3 A Sample Query

Users can launch queries to the UDM through the web user interface provided by OntoFusion as follows. First, the user selects the target concept (TC) and specifies any required attribute filters. Next, the user can add any number of concepts to the query provided that they are related to the TC either directly or indirectly. When adding a new concept to the query, OntoFusion internally checks for this requirement using the unification metadata stored in the UDM. If the selected class fails to be related to the TC, the system rejects to include the concept into the query. Otherwise, the concept is included in the query together with any required attribute filters for that particular concept.

► Figure 3 shows a sample user query involving three concepts. As shown in the figure, the selected TC is the class *Pathology*. The query also includes the attribute filter $\langle \text{Pathology, Name, contains, "cystic fibrosis"} \rangle$ for the selected TC. Additionally, the sample query includes the concepts *Drug and References* – boxed in ► Figure 3. No attribute filters were specified for the classes *Drug* and *References*. The sample query can be read as “retrieve all instances of concept *Pathology* whose name contains the string ‘cystic fibrosis’ and all their related instances of classes *Drug* and *References*”.

Once the query has been executed, the user is presented with the retrieved instances of the concept *Pathology* coming from different sources: RDClinic, OMIM and Orphanet. As depicted in ► Figure 4, each retrieved instance is automatically as-

signed an instance identifier. The latter is composed of the concatenation of the database name – i.e. the database from which the target instance has been retrieved – and the original instance identifier (e.g. the actual value of the instances’ primary key for the RD_clinic database, or the PMID for instances retrieved from PubMed).

If the user selects an instance coming from a structured source, the associated record is shown on the right-hand side of the results window. For instance, as shown in the figure, the user selected the RDClinic_320 instance from the RDClinic database. The available data associated with this instance is shown to the right of the taxonomy of results. Conversely, if the user selected an instance coming from a text-based source – e.g. OMIM or Orphanet – then she would get the actual webpage retrieved from the original source. ► Figure 4 shows two sample instances of the class *Pathology* retrieved from text-based sources – one from OMIM and another one from Orphanet.

Note that it is also possible to retrieve the instances of concepts *Drug* and *References* related to the currently selected instance. As shown in the figure, it would be possible to browse the instances of the concepts *Drug* and *References* related to the RDClinic_320 instance simply by clicking the corresponding button.

Using the provided interface it is possible to query structured and text-based sources in a simple and intuitive manner. The structure of records coming from relational sources is modified to match the schema associated with the target UDM

thus facilitating information fusion. Conversely, the documents – i.e. webpages – retrieved from text-based sources remain unmodified. In this case, the information fusion is implemented by classifying the documents into different categories corresponding to the concepts involved in the query.

As stated before, we tested the correctness of the integration experiment using a test set of 500 queries. The system successfully provided correct results for all queries in the test set.

4.4 Experts’ Feedback

We requested the panel of experts to provide feedback on the results of the integration experiment by answering a questionnaire. The most relevant issues analyzed in the survey include: i) quality and completeness of the extracted schemas, ii) performance of the proposed schema extraction method (in terms of time), iii) query processing performance (in terms of time), iv) completeness of the UDM, v) relevance of the retrieved information, vi) importance of the integration problem, and vii) suitability of the proposed solution. Answers to the questionnaire were provided in a discrete scale (5 = excellent, 4 = good, 3 = average, 2 = fair, 1 = poor). ► Table 6 shows the answers of the experts to the most relevant questions included in the questionnaire.

5. Discussion

The integration of structured and text-based sources is a crucial problem in the biomedical domain. For instance, in the context of the European Commission-funded ACGT [42] project – aimed at setting up a framework to manage clinico-genomic clinical trials in cancer, partnered by the authors – we had to develop methods and tools to combine relational clinical databases developed by hospitals and other health-related institutions with public online textual databases containing genomic data [43].

However, most available biomedical database integration systems do not pro-

Table 6 Summary of results of the questionnaire

	Expert #1	Expert #2	Expert #3	Expert #4
Schema quality	4	4	5	4
Schema completeness	3	4	4	4
Performance of the schema acquisition method	4	4	4	5
Completeness of the UDM	4	5	5	5
Query performance	3	4	3	3
Relevance of retrieved information	3	4	4	4
Importance of the problem	5	5	5	5
Suitability of the proposed solution	4	4	5	4

vide simultaneous support for both structured and text-based sources. On the other hand, systems that provide such functionalities – e.g. TSIMMIS [14] or the updated version of OntoFusion [21] – either i) require the users to have a thorough knowledge of the domain to be able to compose the queries or ii) do not integrate the real text-based sources, but relational copies of the public online resources. The latter issue entails additional problems, such as the automatic updating of the relational copies of the public online databases.

We previously reported [10] a very preliminary smaller integration experiment, for academic purposes, using the methodology described in this paper. In that experiment, we created three textual repositories each including 50 cancer-related documents from PubMed, OMIM and PDB respectively. Then, we integrated such repositories – instead of the actual public online sources – with two relational databases using the proposed method.

Regarding the query processing capabilities of our system, it only uses the basic search facilities provided by the public online sources. It does not take advantage of advanced search functionalities like these provided by PubMed or OMIM. To support these advanced searches it would be necessary to design specialized wrappers to map conceptual objects to specific search fields provided by the sources.

Regarding the scalability of our approach, note that documents belonging to the same public online source are similar in structure – e.g. most documents belonging to the OMIM database contain the subsections *Description*, *Mapping*, *Gene Function*, etc. As the generated schema already includes these generic concepts, the inclusion of additional documents presenting the same internal organization and describing new genetic diseases into the actual source would neither trigger the extraction of a new logical schema nor a re-mapping process. This is another key difference with the experiment reported in [10], in which all extracted concepts – both generic and specific – were used to build a searchable document index to search the document collections using a retrieval engine based on the well known vector-space model [44].

Therefore, the inclusion of new documents describing other genetic diseases would trigger both the generation of a new logical schema and a re-mapping activity.

The integration method presented in this paper can be also regarded as an enabling technique for wider and richer retrieval from multiple and heterogeneous information sources. For instance, in the context of the European ACGT project [42] we plan to use our method to integrate and retrieve clinical records enriched with personalized genomic information collected from public online databases. The latter are not limited to textual resources, since our approach can be adapted to provide support for annotated collections of multimedia objects.

The proposed method is also domain-independent. Thus, it can be used in any context requiring the integration of textual information and relational data.

6. Conclusions

The results of the integration experiment suggest the feasibility of integrating relational and text-based biomedical sources using i) a method to automatically build a logical schema associated with a text-based source, and ii) the methods and tools provided by a readily available structured database integration engine.

By using this approach, text-based sources, together with their extracted schemas, can be regarded as being equivalent to structured sources. Thus, readily available methods and tools for structured database integration such as OntoFusion can be reused to bring together structured and text-based sources. This approach facilitates the integration of differently structured sources, saves time and effort, and benefits from integration methods that have been extensively used and thoroughly tested.

Acknowledgment

This research has been supported by the European Commission through the Advancing Clinico-Genomic Trials on Cancer (IST-2005-026996) and ACTION-Grid projects, the Spanish Ministry of Innovation and Science (RETICS), the Spanish

Ministry of Education (OntoMineBase project, reference TSI2006-13021-C02-01), and the Comunidad de Madrid, Spain. We also thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Sander C. Genomic medicine and the future of health care. *Science* 2001; 287 (5460): 1977–2178.
2. Maojo V, Kulikowski CA. Bioinformatics and Medical Informatics: Collaborations on the Road to Genomic Medicine? *J Am Med Inform Assoc* 2003; 10 (6): 515–522.
3. Knaup P, Ammenwerth E, Brandner R, Brigl B, Fischer G, Garde S, Lang E, Pilgram R, Ruderich F, Singer R, Wolff AC, Haux R, Kulikowski C. Towards Clinical Bioinformatics: Advancing Genomic Medicine with Informatics Methods and Tools. *Methods Inf Med* 2004; 43 (3): 302–307.
4. Martin-Sanchez F, Maojo V, Lopez-Campos G. Integrating Genomics into Health Information Systems. *Methods Inf Med* 2002; 41 (1): 25–30.
5. Maojo V, García-Remesal M, Billhardt H, Alonso-Calvo R, Perez-Rey D, Martin-Sanchez F. Designing new Methodologies for Integration Biomedical Information in Clinical Trials. *Methods Inf Med* 2006; 45 (2): 180–185.
6. Sax U, Schmidt S. Integration of Genomic Data in Electronic Health Records – opportunities and dilemmas. *Methods Inf Med* 2005; 44 (4): 546–550.
7. INFOGENMED: A virtual laboratory for accessing and integrating genetic and medical information for health applications. EC funded project IST-2001-39013.
8. Pérez-Rey D, Maojo V, García-Remesal M, et al. OntoFusion: Ontology Based Integration of Genomic and Clinical Databases. *Comput Biol Med* 2006; 36 (7–8): 712–730.
9. Wiederhold G. Mediators in the Architecture of Future Information Systems. *Computer* 1992; 25 (3): 38–49.
10. García-Remesal M, Maojo V, Crespo J, Billhardt H. Logical Schema Acquisition from Text-Based Sources for Structured and Non-Structured Biomedical Sources Integration. *Proc AMIA Symp* 2007. pp 259–263.
11. Sujansky J. Heterogeneous Database Integration in Biomedicine. *J Biomed Inform* 2001; 34 (4): 285–298.
12. Pyle D. *Business Modeling and Data Mining*. Morgan-Kaufman; 2003.
13. Kersy PJ, Morris L, Hermjakob H, Apweiler R. Integr8: Enhanced Inter-Operability of European Molecular Biology Databases. *Methods Inf Med* 2003; 42 (2): 154–160.
14. García-Molina H, Hammer J, Ireland K, Papakonstantinou Y, Ullman J, Windorn J. Integrating and Accessing Heterogeneous Information Sources in TSIMMIS. *Proceedings of the AAAI Symposium on Information Gathering* 1995. pp 61–64.
15. Mena E, Illarramendi A, Kashyap V, Sheth A. OBSERVER: An approach for query processing in global information systems based on interoperation between pre-existing ontologies. *Distrib Parallel Dat* 2000; 8(2): 223–271.

16. Wache H, Scholz T, Stieghahn H, König-Ries, B. An integration method for the specification of rule-oriented mediators. Proceedings of the International Symposium on Database Applications in Non-Traditional Environments (EFIS 99), Kühlungsborn, Germany, 1999.
17. Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC. Discoverylink: a system for integrated access to life sciences data sources. *IBM Syst J* 2001; 40 (2): 489–511.
18. Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics* 2000; 16 (2): 184–186.
19. Ben Miled Z, Li N, Bukhres O. BACIS: Biological and Chemical Information Integration Systems. *J Database Manage* 2005; 16 (3): 73–85.
20. Kawazoe Y, Ohe K. An Ontology-based Mediator of Clinical Information for Decision Support Systems: a Prototype of a Clinical Alert System for Prescription. *Methods Inf Med* 2008; 47 (6): 549–559.
21. Alonso-Calvo R, Maojo V, Billhardt H, Martin-Sanchez F, García-Remesal M, Pérez-Rey D. An agent- and ontology-based system for integrating public gene, protein, and disease databases. *J Biomed Inform* 2007; 40 (1): 17–29.
22. Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proc AMIA Symp* 2001. pp 17–21.
23. Mason O, Tufis D. Tagging Romanian Texts: a case study for QTAG, a language independent probabilistic tagger. Proceedings of the First International Conference on Language Resources and Evaluation 2000. pp 589–596.
24. <http://www.english.bham.ac.uk/staff/omason/software/qtag.html>. Last accessed: Jan 2008.
25. Woods W. Transition Network Grammars for Natural Language Analysis. *Commun ACM* 1970; 13 (10): 591–606.
26. Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Res* 2004; 32: D267–D270.
27. <http://umlsk.nlm.nih.gov>. Last accessed: Jan 2008.
28. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet* 2000; 25: 25–29.
29. <http://www.geneontology.org/>. Last accessed: Jan 2008.
30. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. The HGNC Database in 2008: A Resource for the Human Genome. *Nucleic Acids Res* 2007. Epub ahead of print.
31. <http://www.genenames.org/>. Last accessed: Jan 2008.
32. Rosse C, Mejino JVL. A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy. *J Biomed Inform* 2003; 36: 478–500.
33. <http://sig.biostr.washington.edu/projects/fm/AboutFM.html>. Last accessed: Jan 2008.
34. Hersh WR, Dickham DH. A comparison of two methods for indexing and retrieval from a full text medical database. *Med Decis Making* 1993; 13 (3): 220–226.
35. Hearst M. Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the 14th Conference on Computational Linguistics 1992. pp 539–545.
36. Friedman E. *Jess in Action: Java Rule-Based Systems*. Greenwich, CT: Manning Publications Co.; 2003.
37. Forgy CL. Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem. *Artif Intell* 1982; 19(1): 17–37.
38. Sinclair J. *Corpus, Concordance, Collocation*. Edinburgh, UK: Oxford University Press; 2000.
39. Gosset WS. The Probable Error of a Mean. *Biometrika* 1908; 6 (1): 1–25.
40. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003; 36: 462–477.
41. Lussier YA, Borlawski T, Rappaport D, Liu Y, Friedman C. PhenoGO: assigning phenotypic context to Gene Ontology annotations with natural language processing. *Pac Symp Bio* 2006. pp 64–75.
42. ACGT – Advancing Clinico Genomic Trials on Cancer. EC funded project FP6-2005-IST-026996.
43. Maojo V, Crespo J, de la Calle G, Barreiro J, García-Remesal M. Using web services for linking genomic data to medical information Systems. *Methods Inf Med* 2007; 46 (4): 484–492.
44. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM* 1975; 18 (11): 613–620.