SAT & ZB: Novel Tools to Acquire and Browse Conceptual Schemas from Public Online Databases for Biomedical Applications

Miguel García-Remesal[1], **Pedro Gil**[1], **Víctor Maojo**[1], **Holger Billhardt**[2] **and José Crespo**[1]

[1] Biomedical Informatics Group
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo S/N, 28660 Boadilla del Monte, Madrid (Spain)
[2]Artificial Intelligence Group
Escuela Técnica Superior de Ingenieros en Informática, Universidad Rey Juan Carlos
C/ Tulipán S/N, 28933 Móstoles (Madrid)

mgarcia@infomed.dia.fi.upm.es

Abstract

In this paper we present a suite of tools to automatically acquire and browse conceptual schemas from large collections of HTML-based biomedical documents. This suite is composed of two tools: the schema acquisition tool (SAT) and the zoomable browser (ZB). The SAT is the implementation of a novel four-phased method to extract conceptual schemas from non-structured sources. First, all documents in the collection are analyzed to extract relevant concepts. Second, the vocabulary discovered during the first phase is organized into a hierarchical structure. Third, the schema is enriched with non-hierarchical *ad-hoc* relationships. The last phase is an optional refinement activity that must be conducted by experts in the domain covered by the collection. The extracted schemas can be navigated using the ZB. We have used these tools for different purposes in the EC funded biomedical research project *Advancing Clinico-Genomic Trials on Cancer* (ACGT), obtaining promising results.

Keywords: Conceptual schema acquisition, concept discovery, hierarchical organization of vocabulary, relationships discovery, schema visualization.

1 Introduction

During the last years, the biomedical community has shown a growing interest in public web-based biomedical databases. Resources such as PUBMED, GENBANK, or OMIM are nowadays being used by biomedical researchers on a daily basis, and also, though less frequently, by health practitioners. For instance, the PUBMED system was queried over 900 million times during 2006, proving the importance of such resources.

Although often presented as keyword-based search engines, public web-based sources are mere web applications built upon large relational databases whose conceptual schemas are not available to researchers. This is a serious drawback, since conceptual schemas provide detailed descriptions of the domain covered by the sources, and thus they are crucial to understand the sources' structure and contents.

In this paper we present a suite of tools to automatically acquire and navigate conceptual schemas from public web-based sources. This suite includes two tools: the schema acquisition tool (SAT), and the zoomable browser (ZB). The SAT is the implementation of a schema extraction method which is a key component of a heterogeneous database integration system developed by the authors called ONTOFUSION (Pérez-Rey *et al.,* 2006). The generated model can be navigated using the ZB. Models extracted by our tool are useful for a wide range of purposes in the biomedical field. In this paper, we describe our experiences using these tools in the context of the EC funded Advancing Clinico-Genomic Trials on Cancer (ACGT) project (The ACGT Consortium, 2005). In ACGT, we used our tools to generate models from several cancer related HTML-based sources whose documents were borrowed from three online databases, namely PUBMED, OMIM, and PDB. We used these models for two different tasks: heterogeneous database integration, and information retrieval, obtaining promising results.

2 Methods

The first tool presented in this paper (SAT) is the implementation of a novel method to mine a conceptual schema describing the domain covered by a collection of HTML-based documents. As shown in figure 1, this method is composed of four sequential phases. During the first three activities, carried out automatically by our tool, HTML documents are examined to discover relevant components of the conceptual model. These components include concepts, hierarchical relationships, and *ad-hoc* relationships. Concepts represent entities or classes of objects belonging to the domain. Hierarchical relationships are asymmetrical relationships targeted to define a hierarchy of concepts. These relationships

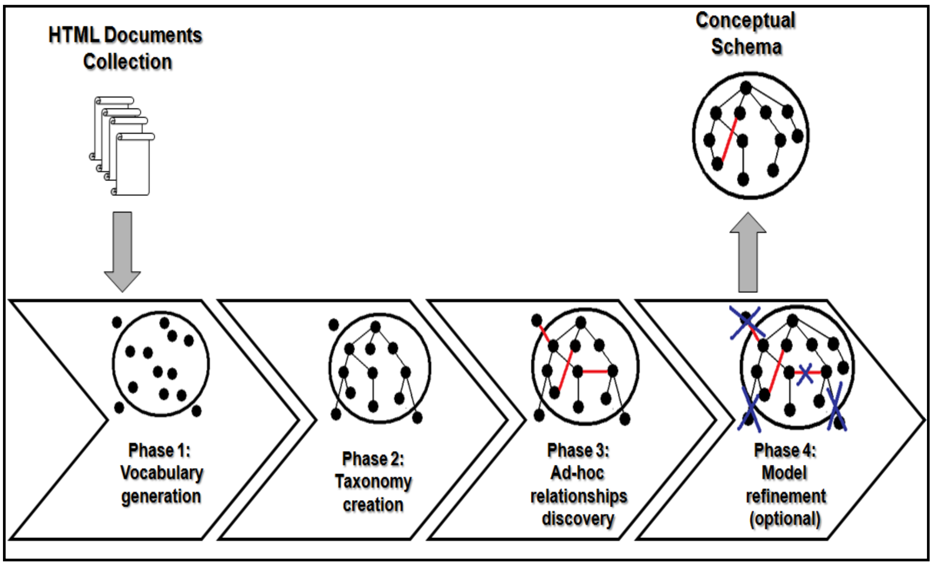

HTML Documents Collection
Conceptual Schema
Phase 1: Vocabulary generation
Phase 2: Taxonomy creation
Phase 3: Ad-hoc relationships discovery
Phase 4: Model refinement (optional)

Fig. 1. Overview of the conceptual schema acquisition method

involve two concepts: the origin and the destination. The origin is called the immediate superior (or super-class) while the destination is called the immediate subordinate (or sub-class). On the other hand, a*d-hoc* relationships are non-hierarchical relations that hold between pairs of concepts. An example of such relationships would be the relation *Tumor_Sample <<sampled-from>> Patient*. Conversely to the previous phases, the fourth phase is an optional refinement activity. In the next paragraphs we briefly describe all the phases of our method.

The first activity is targeted to extract all relevant concepts or noun phrases (NPs) from the text contained in HTLM documents. NPs can be composed of one or more words. To carry out this task, we have designed an algorithm that combines several classic natural language processing (NLP) techniques. Documents are processed as follows.

First, HTML documents are transformed into plain-text documents. To achieve this task, we use a custom parser to detect and remove all HTML-related tags. This parser is also capable of detecting tags and code fragments written using all major web-based scripting and programming languages. Once we have plain text-based documents, these are divided into sentences using a sentence generator. Each sentence corresponding to a given document is fed to a lexicographical analyzer that outputs a set of sorted *tokens* or words. Next, each tokenized phrase is analyzed by a probabilistic part-of-speech (POS) tagger which assigns each token a probability mass function over a predefined set of POS tags. After that, each token is labeled with the most likely POS tag according to its corresponding probability mass function. Finally, the maximum likelihood POS tag sequence associated to each phrase is processed by a transition network manager. The latter includes three distinct transition networks to discover three different types of NPs. This includes simple NPs, conjunctive NPs, and adverbial NPs. Simple NPs are composed of one or more nouns preceded by a succession of adjectives. Similarly, conjunctive NPs can be defined as conjunctions or disjunctions of adjectives followed by a sequence of one or more nouns. By contrast, adverbial NPs are composed of an adverbial form followed by a sequence of adjectives followed by one or more nouns. NPs such as *"water-soluble single-crystalline* TiO_2 *nanoparticles"*, *"cardiovascular and cancer-related mortality"*, and *"freely available peak alignment methods"* are examples of simple, conjunctive, and adverbial NPs that can be discovered by our algorithm. The procedure described above must be iterated N times, being N the number of HTML documents in the collection. During this process, we collect and record some statistics regarding the extracted NPs such as their frequencies or *Term-Frequency* × *Inverse Document Frequency* (TF-IDF) weights (Salton *et al.,* 1988).

To evaluate the relevance of all generated NPs, we combine the use of the recorded statistics and the utilization of a distributed medical and genetic vocabulary server. First, all NPs not meeting a set of predefined criteria based on the statistics are discarded. Second, the remaining NPs are validated against the terminology server. If a given NP cannot be found in the vocabulary server, it is marked as non-relevant, and thus it is a candidate for removal during the refinement phase. Conversely, if the NP has been found in the vocabulary server, the NP is assigned its preferred string—if available—according to the vocabulary server. The latter, powered by ONTOFUSION, provides an integrated means to query three widely accepted biomedical vocabularies and ontologies, namely the Unified Medical Language System (UMLS), the Gene Ontology (GO), and the Human Gene Nomenclature (HGNC). At this moment, the vocabulary server only integrates these three terminologies. However, we are currently working toward the enhancement of our vocabulary server with new resources. In the context of the ongoing ACGT project, we plan to update the server with several cancer-related resources. This includes the National Cancer Institute (NCI) Thesaurus and the ACGT Master Ontology on Cancer, developed by the ACGT Consortium. The NCI Thesaurus, which is aimed to facilitate translational research, is a biomedical vocabulary for cancer research covering terminology across a wide range of cancer research domains. By contrast, the ACGT Master Ontology (The ACGT Consortium, 2005) is a resource targeted to deal with multilevel clinical and genomic data in post-genomic clinical trials. Once the vocabulary has been generated, we can proceed to explain the taxonomical organization activity.

The second activity is aimed to classify all NPs generated during the previous phase into a hierarchical structure. In this work, we have focused our research on finding hierarchical hyponymy relationships—i.e. relationships among a generic concept and its related more specialized concepts. To discover such relationships, we adopted a *pattern-matching* approach. We created a knowledge base containing more than 70 rules or patterns of hyponymy extracted from texts borrowed from PUBMED. To carry out the hierarchical organization of the vocabulary, we proceed as follows. For each NP belonging to the vocabulary, the algorithm collects all sentences from the documents where the target NP occurs. Next, all recorded sentences are analyzed by a rule-based inference engine to find hyponyms of the target concept using the rules stored in the knowledge base. We also included in the knowledge base some heuristics or "rules of thumb" such as the following: *"**IF** a NP is composed of N tokens, with $N \geq 2$ **THEN** the last N-1 tokens constitute an hypernym of the original NP"*. Conversely to the other rules stored in the knowledge base, heuristic rules are not applied to documents, but directly to NPs. For instance, the sample heuristic rule showed above once iterated over the NP "*water-soluble single-crystalline TiO_2 nanoparticles"* outputs the following hierarchical relations: *"nanoparticles"* ← *"TiO_2 nanoparticles"* ← *"single-crystalline TiO_2 nanoparticles"* ← "*water-soluble single-crystalline TiO_2 nanoparticles"*, where *(b ← a)* denotes the hyponymy relationship *"a is a sub-concept of b"*.

The result of this activity is a directed graph that represents a hierarchy of NPs. Every node in this hierarchy—except the root node—must be related to its parent node by a hyponymy relationship. After the execution of the algorithm, all orphan nodes, if any, are automatically assigned the root node as parent. Once the taxonomy has been created, we can proceed to the third activity.

The goal of the third phase is to enrich the hierarchy of NPs generated during the previous phases with *ad-hoc* relationships. The latter are links between pairs of concepts that denote non-hierarchical relationships. A simple example of an *ad-hoc* relationship would be *codon* ← *<encodes>, <is_encoded_by>* → *aminoacid*. This *ad-hoc* relationship states that a codon *encodes* an aminoacid and that an aminoacid *is encoded* by a codon. The labels <*encodes*> and <*is_encoded_by*> are the *roles* played by concepts *codon* and *aminoacid* in the relationship. We must clarify that the method we outline in this section to discover *ad-hoc* relationships does not extract role names. This method is based solely on co-occurrence statistics, and thus it is only useful to discover relationships between concepts, not for discovering role names. Nevertheless, we are currently working on an hybrid approach that combines statistics and NLP techniques to extract role names.

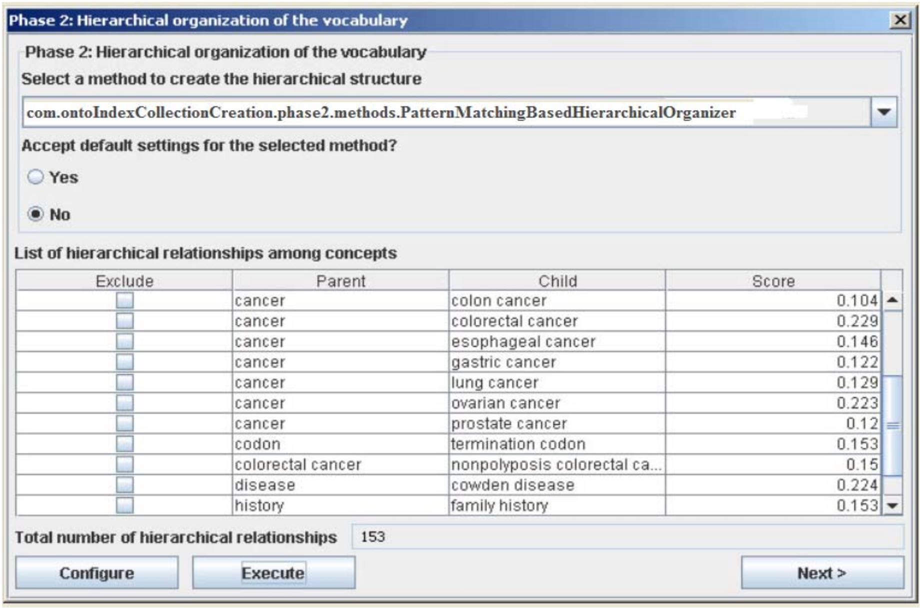
To discover *ad-hoc* relationships, we perform an hypothesis testing under the following null hypothesis: *"The occurrence of concept A is independent of the occurrence of concept B in the same context"*. Thus, to conduct the hypothesis testing, we compare the expected and observed co-occurrence frequencies between concepts *A* and *B*. If statistically significant differences are observed, we can reject the null hypothesis. This situation indicates that the occurrence of concept *A* is conditioned on the occurrence of concept *B* in the same context. This can be considered as preliminary evidence that a semantic relationship holds between concepts *A* and *B*. Observed frequencies are computed in small contexts called *concordances.* Concordances are $2s+1$ sized sequences of NPs centered in a given concept which is called the *node concept*. By centered, we mean that the node concept is located in the $(2s/2) + 1$ position of the sequence. Typical values for s are between 3 and 15. Concordances can be obtained from document surrogates. A document surrogate is a list of all NPs extracted from a given document sorted in the same order as they appear in the original document. To calculate the observed co-occurrence frequency between two given NPs c_1 and c_2 we proceed as follows. First, we choose one of the NPs, say c_1, as the node concept. Next, we gather all concordances centered at c_1 from the document collection. After that, we calculate the observed frequency of co-occurrence between both NPs as the sum of all occurrences of c_2 over the whole set of concordances. Analogously, expected frequencies are computed from a large set of documents extracted from the MEDLINE database covering the same domain as the actual document collection. The statistic used to conduct the hypothesis testing is the T-Score, based on the Student's T statistic. The T-Score values greater than 2 suggest that the null hypothesis must be rejected, and thus there exists an *ad-hoc* relationship between both NPs. The greater is the value of this statistic, the stronger is the existing relationship between both NPs.

The fourth activity is an optional refinement phase that must be manually conducted by experts in the domain assisted by a knowledge engineer. In this activity, the experts can remove incorrect or irrelevant concepts, hierarchical relationships, and *"ad-hoc"* relationships. All statistics and information recorded during the previous phases, as well as the vocabulary server, are available to curators to assist them during the refinement process.

Once this process has been completed, the outcome is a refined conceptual model that describes the domain covered by the document collection. This conceptual model can be navigated using our zoomable browser. In the next section we describe both the tool we have implemented to support the described four-phased method and the schema browser.

3 Results

Both the SAT and the ZB have been implemented using the Java Programming Language. The SAT is a stand-alone application that can be executed in any platform and/or operating system as long as they provide a recent



Phase 2: Hierarchical organization of the vocabulary

Phase 2: Hierarchical organization of the vocabulary

Select a method to create the hierarchical structure: com.ontoIndexCollectionCreation.phase2.methods.PatternMatchingBasedHierarchicalOrganizer

Accept default settings for the selected method?

○ Yes

● No

List of hierarchical relationships among concepts

Exclude	Parent	Child	Score
□	cancer	colon cancer	0.104
□	cancer	colorectal cancer	0.229
□	cancer	esophageal cancer	0.146
□	cancer	gastric cancer	0.122
□	cancer	lung cancer	0.129
□	cancer	ovarian cancer	0.223
□	cancer	prostate cancer	0.12
□	codon	termination codon	0.153
□	colorectal cancer	nonpolyposis colorectal ca...	0.15
□	disease	cowden disease	0.224
□	history	family history	0.153

Total number of hierarchical relationships: 153

Fig. 2. A screenshot of the SAT after completion of phases 1 and 2 applied to a cancer-related HTML document collection

Java Virtual Machine implementation. Similarly, the ZB can be accessed through the Internet using any Java-enabled web browser.

To use the SAT, the administrator must provide either a local copy of the HTML-based collection, or an URL pointing to the collection location on the Internet. In the latter case, the document collection will be composed of the page associated to the provided link, plus all reachable pages belonging to the same Internet domain as the initial web page.

Once the collection has been preprocessed and loaded, the administrator can proceed to the vocabulary generation phase. Activities 1 through 3 are fully automated and only require user interaction to proceed to the next phase.

Figure 2 shows a screenshot of the SAT after the execution of phases 1 to 2—i.e. vocabulary discovery and hierarchical organization of vocabulary—applied to a cancer-related document collection.

As depicted in figure 2, hierarchical relationships are presented in a tabular view. The information provided by the relationships table includes the parent class or concept, the child class, and other quality measures and statistics—i.e. results of hypothesis testing.

As shown in the table, the vocabulary discovered during phase 1 involves both single word-based NPs—e.g. *cancer*, *disease*, or *history*—and multi-word NPs such as *esophageal cancer*, *Cowden disease*, *or family history*. Besides, the hierarchical relationships extracted by the SAT are coherent and highly representative of the domain of interest. Note that although the SAT does not require any user interaction, the table contains a "exclude" checkbox for each generated relationship. This checkbox can be checked out by the administrator to delete incorrect or irrelevant hierarchical relationships generated by the SAT. This can be done either at this stage or in phase 4—i.e. optional refinement activity. To proceed to the next phase—i.e. *ad-hoc* relationships discovery—the administrator only has to click on the "Next" button.

The algorithms used by SAT in each of the phases are those described in section 2. However, it is also possible to utilize user defined methods and algorithms—i.e. implemented by the user—if required. For instance, in phase 2, we have used the pattern matching-based hierarchical organizer outlined in section 2. However, we could be interested in using a different method such as for instance the Rada algorithm (Forsyth *et al.*, 1992). The only requirement is that the class that implements such algorithm must also implement the interface *HierarchicalOrganizer*.

Regarding the VZ, it was created using the Java Programming Language and the Swing library, and it can be accessed using any Java-enabled browser. Figure 3 depicts a screenshot of the VZ showing the generated model for the cancer-related collection.

As shown in the figure, it is composed of two sub-windows. The leftmost window shows the extracted model represented as a tree. Concepts are represented by yellow circles, while relationships are denoted by green circles. Hierarchical relationships are represented by using indentation—e.g. *cancer* and *breast cancer* or *codon* and *codon termination*. Conversely, the rightmost window renders the model in a more attractive and natural manner. Users only have to drag the screen to browse the model. It is also possible to navigate the relationships just by clicking on the corresponding button. For instance, if we clicked on the button next to the relationship *TO.mutation* associated to the concept *cancer*, the browser would redirect us to the exact location of concept *mutation* within the hierarchy.

The browser has many other interesting navigation features such as class filtering, zooming effects, or rotations, thus being a powerful tool for conceptual schema browsing. Besides, it is also possible to export the

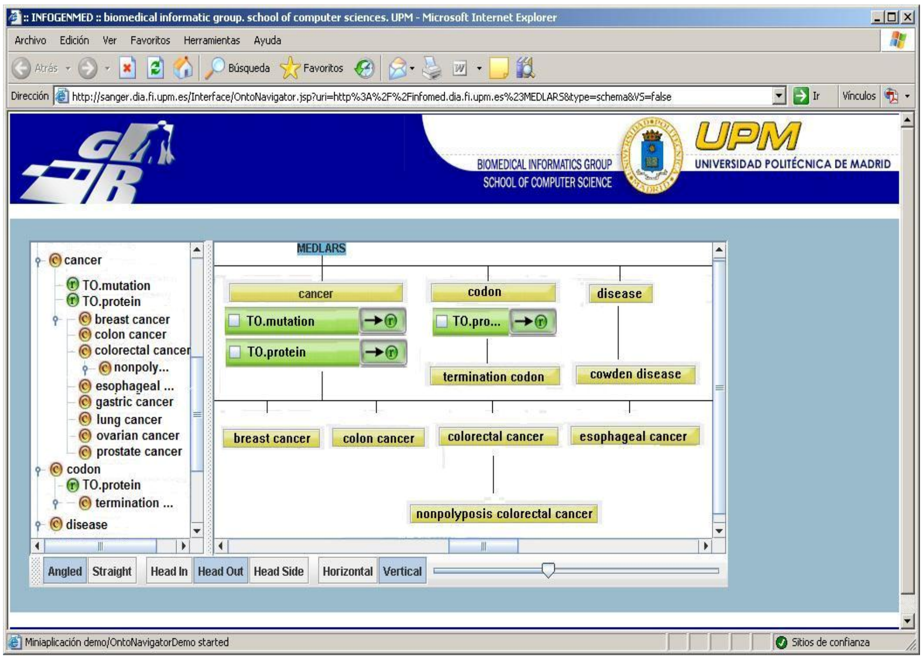

:: INFOGENMED :: biomedical informatic group. school of computer sciences. UPM - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos
Dirección http://sanger.dia.fi.upm.es/Interface/OntoNavigator.jsp?uri=http%3A%2F%2Finfomed.dia.fi.upm.es%23MEDLARS&type=schema&VS=false
Ir Vínculos
BIOMEDICAL INFORMATICS GROUP
SCHOOL OF COMPUTER SCIENCE
UPM
UNIVERSIDAD POLITÉCNICA DE MADRID
cancer
TO.mutation
TO.protein
breast cancer
colon cancer
colorectal cancer
nonpoly...
esophageal ...
gastric cancer
lung cancer
ovarian cancer
prostate cancer
codon
TO.protein
termination ...
disease
MEDLARS
cancer
codon
disease
TO.mutation
TO.pro...
TO.protein
termination codon
cowden disease
breast cancer
colon cancer
colorectal cancer
esophageal cancer
nonpolyposis colorectal cancer
Angled
Straight
Head In
Head Out
Head Side
Horizontal
Vertical
Miniaplicación demo/OntoNavigatorDemo started
Sitios de confianza

Fig. 3. A screenshot of the VZ showing the generated model for the cancer-related document collection

schema to several formats such as RDF, RDF(S), DAML+OIL, or OWL for offline browsing.

Once we have briefly described the tools, in the next section we describe some experiments using SAT and VZ in the context of a biomedical research project, the EC funded ACGT project.

4 Applications of SAT & VZ to biomedical research projects

ACGT (The ACGT Consortium, 2005) is an IST-FP6 integrated project, funded by the European Commission for the development of services to support clinic-genomic trials on cancer in a grid-based environment. In these trials, physicians and researchers need to access heterogeneous and disparate data sources. Semantic access to these data and the possibility to integrate them seamlessly are issues that ACGT aims to solve.

In the context of the ACGT project, one of the main challenges is to achieve the integration of structured and non-structured biomedical sources. Structured sources can be defined as those sources equipped with a logical schema describing their information contents—i.e. the portion of the domain of interest covered by the source. Record-based structures are the main information units in this kind of sources. Examples of structured sources are relational or object-oriented databases. Conversely, non-structured sources can be defined as schemaless sources, where documents are the basic information unit. Text or HTML based document collections are common examples of non-structured sources.

During the development of a previous EC funded project, named INFOGENMED (The INFOGENMED Consortium, 2001), we developed the ONTOFUSION system (Pérez-Rey *et al.*, 2006), which provides methods and tools to integrate structured sources. In the context of ONTOFUSION, sources are represented by *domain models* (DMs). DMs are common conceptual representations that describe the portion of the domain covered by a given source. DMs are created by mapping objects from the sources' logical schema to objects belonging to a global domain model (GDM). The latter contains all the needed objects named with normalized terminology. Once a DM has been created for all sources to be bridged together, a unification process is performed. The latter produces a unified DM that covers all the information stored in the underlying sources.

Unfortunately, methods and tools provided by ONTOFUSION cannot be reused to integrate structured and non-structured sources. Mapping and unification processes are useless when dealing with non-structured sources, since they lack a logical schema.

In the context of an experiment carried out for the ACGT project, we had to integrate a set of five cancer-related sources. Two of them were structured—relational—databases, and the rest collections of HTML documents borrowed from three public online biomedical databases, namely PUBMED, OMIM, and PDB. To solve the abovementioned problem, we used SAT to generate a conceptual schema for each of the non-structured sources. Once we equipped each of the non-structured sources with a logical schema, we were enabled to use the methods and tools provided by ONTOFUSION—mapping and unification—to seamlessly integrate the sources. Figure 4 shows an extract of the model generated by SAT for the PDB-based non-structured source.

PDB (or Protein Data Bank), is a database that describes the tree-dimensional structure of a protein or nucleic acid, as determined by X-ray crystallography or nuclear magnetic resonance (NMR) imaging (Kouranov *et al.*, 2006). As can be seen, concepts and relationships automatically acquired by our tool are coherent and very representative of the domain of interest. It includes concepts such as *cancer*, *tumor*, *protein*, *structure*, or *DNA* as well as all important relationships between pairs of concepts.

After the execution of the mapping and unification processes, we obtained a DM that represents the whole information space covered by the 5 sources. This DM was composed of 257 concepts, 106 hierarchical relationships, and 425 *ad-hoc* relationships. The integrated source has been successfully used in a large number of tasks related to the ACGT project, proving the validity of our approach.

On the other hand, we also have used the SAT to improve keyword-based document retrieval performance in large collections of cancer-related documents. We have developed a novel information retrieval model known as OIM (*Ontological Indexing Method*) that exploits the domain knowledge contained in the conceptual schemas generated by the SAT to expand and improve queries. Our model outperformed the gold standard—the vector space model (Salton *et al.*, 1975)—for three cancer-related test collections.

5 Conclusions

We have created a suite of tools (SAT & ZB) to automatically acquire and visualize conceptual schemas from public online biomedical databases. The SAT generates highly coherent schemas that describe the domain covered by the target sources. On the other hand, the ZB allows users to navigate the acquired schemas and even to export them to different formats. We have used our tools in the context of the EC funded ACGT project for two different tasks: heterogeneous cancer-related data sources integration and document retrieval. Our experience using these tools in ACGT has been very positive, thus proving their utility on biomedical research projects involving public online databases.

6 Acknowledgements

This research has been supported by the Advancing Clinico-Genomic Trials in Cancer (ACGT) project (IST-2005-026996) funded by the European Commission, and the OntoMineBase project (TSI 2006-13021-C02-01), funded by the Spanish Ministry of Education and Science.

7 References

Pérez-Rey, D., Maojo, V., García-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martín-Sánchez, F., Sousa, A. (2006): ONTOFUSION: Ontology-based Integration of Genomic and Clinical Databases. *Computers in Biology and Medicine* **36**(7-8):712-30.

The ACGT Consortium (2005): ACGT: Advancing Clinico-Genomic Trials on Cancer. EC funded project IST-2005-026996.

Salton, G., Buckley, C. (1988): Term-weighting approaches in automatic text retrieval. Information *Processing & Management* **24**(5): 513–523.

Forsyth, R., Rada, R. (1992): *Machine Learning: Applications in Expert Systems and Information Retrieval*. Halsted Press, New York, NY, USA.

The INFOGENMED Consortium (2001): INFOGENMED: A virtual laboratory for accessing and integrating genetic and medical information for health applications. EC funded project IST-2001-39013.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E., Berman, H.M. (2006): The RCSB PDB information portal for structural genomics. *Nucleic Acids Research* **34**, Database issue D302-D305.

Salton G., Wong A., Yang C.S. (1975): A Vector Space Model for Automatic Indexing. *Communications of the ACM* **18**(11):613-20.