Ontology-Based Semantic Annotations for Biochip Domain

Khaled Khelif[1], Rose Dieng-Kuntz[1]

[1] INRIA, UR Sophia Antipolis project ACACIA
2004, route des lucioles BP93, 06902 Sophia Antipolis Cedex, France
{khaled.khelif, rose.dieng}@sophia.inria.fr

Abstract : We propose a semi-automatic method using the information extraction (IE) techniques for facilitating the generation of ontology-based annotations for scientific articles. Furthermore, we evaluate and discuss our method by applying it to the annotation of textual corpus provided by biologists working in the biochip domain. Finally, we argue that ontology-based semantic annotation can improve information retrieval.

1 Introduction

The documents published on the Web represent a very important source of knowledge. This knowledge is essential for checking, validating and enriching of a research work. It is the case of research in the domain of molecular biology and more particularly in the domain of DNA-microarray experiments.

These DNA-microarray experiments that can assess tens of thousands of genes simultaneously provide a huge amount of information: for example, information about the roles played by particular genes in drug sensitivity, the effects of drugs on gene expression and the effects of genetic mutations on sensitivity and response [24].

These experiments present difficulties for a biologist, in particular when s/he validates and interprets the obtained results. First, using a classic search engine, s/he has to search in document bases or genetic data bases using keywords corresponding to genes and the biological phenomenon studied, documents which argue, confirm or invalidate his/her hypotheses, then s/he must analyse the documents found in order to identify relevant knowledge.

Semantic Web [2] techniques can facilitate this task of information retrieval: it can be carried out by associating to each document a semantic annotation based on an ontology describing the domain. This annotation will then describe the semantic contents of the documents. In the case of DNA-microarray experiments, the relevant information is: the type of genes intervening in the experiment described by the article and the interactions that can exist between these genes, and cellular components or biological processes.

In spite of its advantages, the creation of a semantic annotation is a difficult and expensive process (time, people...) for the biologists. The automatic information extraction from texts can thus be an alternative for the generation of these annotations.

In the framework of a collaborative project with biologists working on DNA-microarray experiments at IPMC[1] (l'Institut de Pharmacologie Moléculaire et Cellulaire), our work consists of assisting them in their experiments and facilitating their validation and interpretation of the obtained results. Since this phase of validation is based on information retrieval, our approach rests on the semi-automatic generation of semantic annotations for scientific articles in the DNA-microarray domain. These articles can come from internal sources such as specific documentation databases for each biologist or from external sources such as on line documentation databases (e.g. Medline[3]).

Thus, we developed a method which, starting from a text written by a biologist (e.g. scientific articles), allows generating a structured semantic annotation, based on a domain ontology, and describing the semantic contents of this text. These annotations can, in turn, be used for retrieving the relevant documents for biologists who want to validate their experimental results.

[1] http://www.ipmc.cnrs.fr/

2 Background work

2.1 Ontology-based semantic annotation

The goal of the Semantic Web is to structure the contents of the Web and allow machines to process these contents and to reason on the knowledge represented in the Web pages in particular to facilitate navigation and information retrieval for humans.

One of the most important layers in the architecture of the semantic Web is hence the knowledge representation layer, as it provides formal grounding for representing the semantics of the information and documents on the web. Formal semantics are commonly embedded in ontologies. Defined by [10] as a specification of a conceptualization, an ontology comprises a set of concepts describing a domain and a set of relationships between these concepts. In the semantic web context, these ontologies are used to annotate the domain resources (persons, documents …).

In addition to simple information such as the title and the authors, a "semantic" annotation provides a more precise description of the knowledge contained in the document and its semantics in the domain. A semantic annotation must be well defined, easy to understand by the domain experts and not ambiguous. To fulfill these requirements, a semantic annotation should be based on a formal model of the domain.

The formalization of the annotation scheme using the ontological hierarchy, enables annotators to choose the appropriate level of annotation detail, helps to constrain its structure, to diminish ambiguity and to reduce errors in the annotation process.

In addition, the fact that annotation is based on an ontology leads to use standard formalisms such as RDF [13] or OWL [15] which allow the reuse of these annotations by different annotation tools and search engines.

2.2 Ontologies in the biomedical domain

In medicine and biology, exhaustive domain ontologies have been developed and are constantly incorporating new pieces of knowledge. Among them, let us cite, the Gene Ontology [1], Galen[2] (General Architecture for Language and Nomenclatures), Menelas [25].

These ontologies are used as the initial knowledge base of semantic retrieval system and provide a good basis for the development of Semantic Web applications for medicine purposes.

To facilitate the navigation and the use of these different sources, the U.S. National Library of Medicine (NLM) initiated in 1986 the Unified Medical Language System (UMLS) project. Its goal is to help health professionals and researchers to use biomedical information from a variety of different sources [14]. The UMLS consists of:

- UMLS Metathesaurus: This repository collects millions of terms belonging to the most important nomenclatures and terminologies defined in the biomedical domain.
- UMLS Semantic Network: it consists of 134 Semantic Types with 54 possible links between these types, and it represents a high-level abstraction from the UMLS Metathesaurus [18].

To describe the biochip domain which covers a great part of biomedicine domain (drugs, cells, genes, processes …), we chose UMLS.

We considered the UMLS semantic network as a general ontology for the bio-medical domain: the hierarchy of semantic types can be considered as a hierarchy of concepts and the terms of the metathesaurus can be considered as instances of these concepts.

[2] Ontology GALEN. http://www.opengalen.org, 2001

3 Description of our method

3.1 Motivations

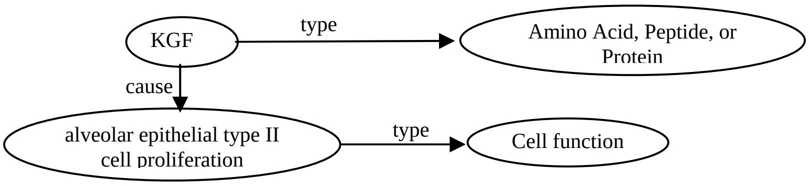
Since our goal is to facilitate the information retrieval task for biologists, we wanted to know in which information the biologist would be interested in an article in order to be able to create relevant annotations. We thus studied how a biologist annotates a document, so we provided three biologists with the same articles and asked them to annotate them manually.

This study revealed several common points between biologists annotations, even if the ways of annotating were different, the information selected by different biologists was almost the same.

We noticed that the biologists underlined primarily the names of genes, substances or proteins studied, the biological phenomenon or the cellular functions treated as well as the verbs which describe a relation between these various elements.

Example of sentence annotated by the three biologists: “KGF causes alveolar epithelial type II cell proliferation”, this sentence identifies that a substance (KGF), is related to a certain type of cellular function (cell proliferation) with a causal relation (causes).

By using the UMLS semantic network, this annotation can be represented by the graph below and translated towards any knowledge description formalism.

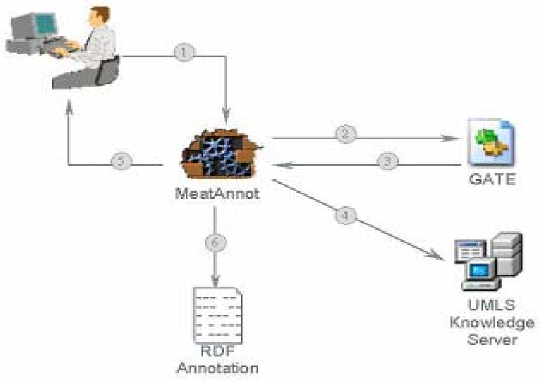

KGF
type
Amino Acid, Peptide, or Protein
cause
alveolar epithelial type II cell proliferation
type
Cell function

Our work aim to automate the extraction of this kind of information and the generation of ontology-based semantic annotation to describe it.

3.2 Our method

We present here a method to generate semi-automatically semantic annotations on articles related to the biochip domain.

The following figure describes the general architecture of the system MeatAnnot built using our method:



1. User gives document to annotate
2. MeatAnnot sends the document to GATE
3. Document is tokenized and postagged by GATE
4. MeatAnnot extracts terms and relations and validates them via UMLSKS
5. MeatAnnot presents extracted information to the user for validation
6. MeatAnnot generates RDF annotation for the document

Figure 1. General architecture of the proposed system

3.2.1 Step 1: Term extraction

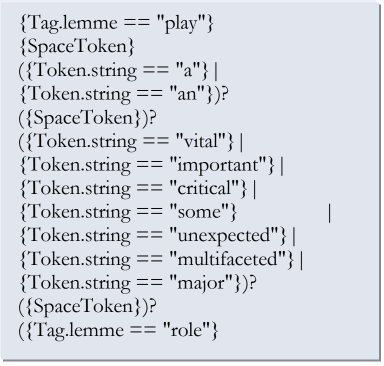
In this step we used two modules of GATE [8] (General Architecture for Text Engineering), the Tokeniser and the Postagger. The tokeniser splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an initial capital, all upper case, etc.), and the Postagger (we used TreeTagger[3]) assigns a part-of-speech tag (verb, noun…) to each word or symbol.
After tokenizing and tagging texts, we used an extraction window of size four (four successive words are considered as a candidate term). For each candidate term, if it exists in UMLS, we process the following word otherwise we decrease the size of the window till zero.
To increase relevance of the extracted terms and decrease the execution time, we considered that a candidate term cannot begin with a verb, a stop-word (preposition, pronoun…) or a symbol.

3.2.2 Step 2: UMLS interrogation

To facilitate UMLS use, the U.S National Library of Medicine (NLM) has created the UMLSKS (KS: Knowledge server). This server provides access and navigation in the metathesaurus and the semantic network of UMLS. We used this server to validate our candidate terms.
For each candidate term extracted, we send a query to UMLSKS, the answer received in XML format (if the term exists in umls) is parsed to obtain information about the term (its semantic type, its synonyms …).

3.2.3 Step 3: Relation extraction

In this step we used Syntex [4] to reveal the "verb syntagms" usually used by the authors of the scientific articles constituting the textual corpus. These verb syntagms can constitute potential relations between bio-medical concepts. Then, we used JAPE [8], a language based on regular expressions and allowing writing information extraction grammar from texts processed by GATE. So, for each relation revealed by Syntex, we created manually an extraction grammar to extract all the instances of this relation as well as the concepts which are linked by this relation.
Example of grammar:



```
{Tag.lemme == "play"}
{SpaceToken}
({Token.string == "a"} |
{Token.string == "an"})?
({SpaceToken})?
({Token.string == "vital"} |
{Token.string == "important"} |
{Token.string == "critical"} |
{Token.string == "some"}              |
{Token.string == "unexpected"} |
{Token.string == "multifaceted"} |
{Token.string == "major"})?
({SpaceToken})?
({Tag.lemme == "role"}
```

This grammar allows detection of instances of the semantic relation "Play a role" with its different lexical forms in the textual corpus.

3.2.4 Step 4: Annotation generation

In the goal of the extension of the web towards semantic web, several formalisms have emerged, like RDF(S), DAML-OIL and OWL to describe ontology and semantic annotations. In our approach, we chose RDF Schema to define the ontology and RDF to edit the annotations. We developed a script which allows automatic translation of the UMLS semantic network from its textual format to an RDFS ontology.
Using this ontology, we developed a graphical interface for collecting information provided by linguistic tools and presented them to user for validation. Then, an RDF annotation describing the validated infor-

[3] http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html.

mation is generated automatically by our tool, associated to the studied article and stored in the directory containing the annotations of the other articles.

The example below summarizes the process steps. Let us consider an article related to lung development and containing the sentence:

> *« HGF plays an important role in lung development »*

Information extracted from this sentence is:

- HGF : instance of the “Amino Acid, Peptide, or Protein” concept of UMLS;
- Lung development : instance of the “Organ or Tissue Function” concept of UMLS;
- HGF play role lung development: instance of “play role” relationship between the two instances ‘HGF’ and ‘lung development’.

The RDF annotation generated is:

```
<rdf:RDF
   xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
   xmlns:m='http://www.inria.fr/acacia/meat#'
   xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'>
<m:Amino_Acid_Peptide_or_Protein rdf:about='HGF#'>
          <m:play_role>
           <m:Organ_or_Tissue_Function rdf:about='lung
            development#'/>
          </m:play_role>
</m:Amino_Acid_Peptide_or_Protein>
</rdf:RDF>
```

4 Validation and discussion

Our test corpus (a set of articles related to lung diseases) was provided by a research team of the IPMC, working on DNA-microarray experiments. Our aim in this test phase was essentially: (i) the validation of the concept instantiation method (sect 4.1), (ii) the validation of the relation extraction method (sect 4.2) and (iii) the verification of the coherence and consistency of the generated annotations (sect 4.3).

4.1 Concepts instantiation

The corpus analysis by our tool Meat Annot enabled us to extract the majority of the UMLS terms from text. The few noticed exceptions are due, in general, to the orthography mistakes made by the authors of the articles, to the use of abbreviations and last, to the use of special characters such as Latin characters.

Extracted terms are automatically linked to their concepts and then used for the generation of the annotation.

4.2 Relation instantiation

In this phase we used the grammars presented in section 3.2.3. For example, for the relation “Play role”, the tool extracted 35 occurrences from the 49 appeared in the textual corpus. Our study of the corpus revealed that these errors are generally due linguistic variations in the text such as, in the sentence “*the key role that endogenous KGF has been shown to play in wound healing in the skin*”.

After extracting the relations, our tool detects terms linked to each instance of the relation to propose the adequate annotation. Three cases are dealt with in this phase:

- Case 1: The terms detected at the left and the right of the relation belong to the UMLS metathesaurus, the annotation is automatically generated by linking each term to its concept and to the instance of the relation (see section 3.2.4).
- Case 2: There is a conjunction of UMLS terms on the left and/or the right of the relation, for example: “*KGF and HGF play role in pulmonary inflammation*”. The tool factorizes UMLS terms and generates annotations for each one: in the previous example, it generates two annotations for the sentences “*KGF play role in pulmonary inflammation*” and “*HGF play role in pulmonary inflammation*”.
- Case 3: there are no UMLS terms on the left and/or the right of the relation. The tool presents the sentence where the relation appeared and the user then indicates the terms linked to this relation (if they exist).

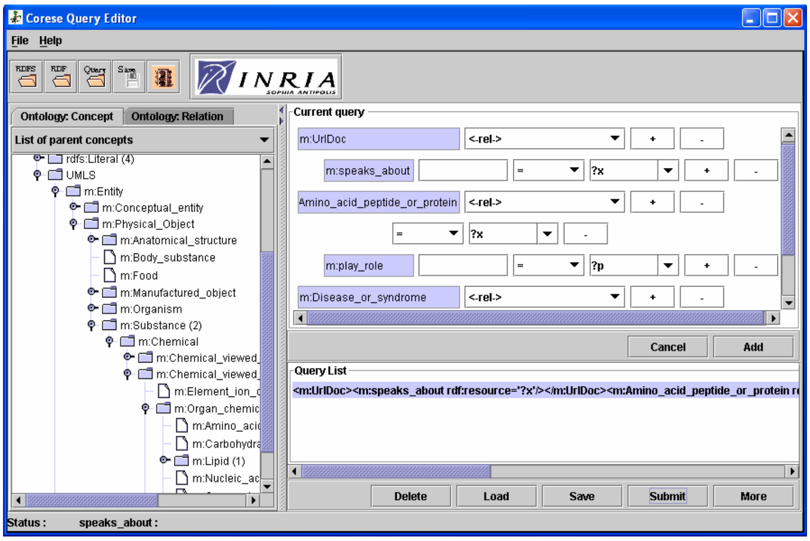
4.3 Coherence and consistency of the annotations

Our aim in this test phase was the verification of the coherence and the consistency of the generated annotation, so we used the semantic search engine CORESE [7]. This engine enables:

- To load ontologies formalized in RDF Schema;
- To load RDF annotations based on the ontology;
- To process queries on the annotations as well as on the ontology.

Basing ourselves on the observations which we have made on the working methods of biologists, we developed a scenario, where a biologist who made a biochip experiment to reveal amino acids (sequences of genes) intervening in the lung diseases, tries to validate the coherence of his results by retrieving articles related to this type of experiments.

Thus, we loaded the UMLS ontology in CORESE and we processed this query on our generated annotations (see Fig2).



<m:UrlDoc><m:speaks_about rdf:resource='?x'/></m:UrlDoc><m:Amino_acid_peptide_or_protein r

Figure 2. CORESE interface showing the ontology and the query structure

Figure 2 shows the concepts hierarchy of the UMLS ontology on the left and a query which enables to find all articles expressing that a particular amino acid plays a role in the lung diseases on the right.

This query requires domain ontology information about a “amino acid” whose name is “x” (any amino acid) and that plays a role in “diseases”. The keyword-based retrieval methods cannot establish such a relation, they can only match the keywords “amino acid”, “role” and “diseases” when the metadata descriptions of articles mention those keywords.

In addition, this example shows that the use of the ontology-based semantic annotations with an engine like CORESE, allows the user to convert a natural language query into a precise semantic schema and thus increase the relevance of his results.

5 RELATED WORK

A number of annotation systems for generating semantic annotation exist. The most interesting of these are Artequakt [12], MnM [23] and OntoMat. A commercial version of OntoMat is available as OntoAnnotate[4].

Artequakt is a project for the generation of personalized narrative biographies of artists from fragments of information extracted from web pages. It has three key components: knowledge extraction, information management, and biography construction. The knowledge extraction module uses GATE to extract named entities and instantiate ontology concepts, after that it submits queries to the ontology to obtain binary relation between extracted terms. To reduce linguistic variation between relations defined in the

[4] http://www.ontoprise.de/products/ontoannotate

ontology and the extracted facts, Artequakt uses lexical chains (synonyms, hypernyms, and hyponyms) defined in WordNet.

MnM and OntoMat are very similar. Both use the Amilcare [6] information extraction system which is designed to support active annotation of documents. To use Amilcare, the user has to manually annotate a training set of documents by ontology concepts. This training set is learned by Amilcare to generate extraction rules which can be used to extract information from text. In contrast with OntoMat, MnM can handle multiple ontologies in the same time. On the other hand OntoMat, stores annotated pages in DAML+OIL using OntoBroker[5]. But, both differ from our system which uses linguistic analyses to extract information and does not support learning techniques.

Among the systems listed above, MeatAnnot is similar to Artequakt in spirit. Both use GATE to extract information and try to identify relations between the concepts. However, while Artequakt extracts only named entities to instantiate concepts, populate ontology and generate biographies, Meat Annot attempts to extract all information concerning experiments (processes, diseases, protocols…) and provides an RDF annotation repository to facilitate information retrieval.

Our method can also be compared with (a) work exploiting information extraction for the biology domain [22] (b) and with work on the generation of semantic annotations for the semantic web [11]. In the domain of article mining in biology, [20] proposes statistical techniques and machine-learning algorithms for discovering interactions among genes from article abstracts in biology in PubMed base. Our approach relying on linguistic tools differ from machine-learning-based approaches proposed by [3] [16].

Concerning automatic generation of RDF annotations, our approach differs from the approach presented in [5] that generalizes structured document annotations from an example of manual annotation. It also differs from [9] that offers generation of annotations consisting of concept instances, in order to enrich an ontology: our approach allows the generation of semantic annotations based not only on concept instances but also on relation instances. Relation extraction was studied by several researchers: in [19], the CAMELEON method which allows the extraction of semantic relations between terms using markers; in [16], after a learning phase on a textual corpus, rules for the extraction of relations between genes and proteins are generated; in [17], the authors use a shallow parsing of local structures around verbs to extract gene-gene and protein-protein interactions.

6 CONCLUSIONS

In this paper, we presented a method to facilitate the generation of semantic annotation using ontology.

The use of ontology-based semantic annotations improves the information retrieval [21]. Our aim is to facilitate the validation and the interpretation of DNA-microarray experiments results by using this method which was approved by IPMC biologists who found the generated annotations relevant.

As a further work, the MeatAnnot prototype will be improved by (a) offering to users the possibility to add new concept instances which do not belong to the UMLS metathesaurus, (b) refining the extraction techniques, (c) developing heuristics which propose terms in the neighborhood of a relation in order to generate new instances not detected automatically.

Finally, this method is generic, it is independent of:

- Domain: it can be applied on any domain as soon as the domain description and needs are available;
- NLP tools: the linguistic analysis is not complex and does not require a specific NLP tool;
- Ontology: it can be based on any domain ontology.

7 Acknowledgment

We thank very much the team of Pascal Barbry from IPMC and for making available their expertise in biochips domain, Didier Bourigault for providing us the results of Syntex on our corpus, Laurent Alamarguy for his assistance in linguistic domain, Remy Bars from Bayers Cropscience for fruitful discussions and PACA[6] region which finances this work by a regional grant.

[5] http://www.ontoprise.de/products/ontobroker

[6] http://www.cr-paca.fr/

References

1. Ashburner M. et al.,"Gene ontology: tool for the unification of biology. The Gene Ontology Consortium" Nat. Genet. 9-25. (2000)
2. Berners-Lee T., Hendler J. & Lassila O., The Semantic Web, Scientific American, 84(5) p. 34-43. (2001)
3. Blaschke C. & Valencia A., Molecular biology nomenclature thwarts information-extarction progress. IEEE Intelligent Systems & their Applications, 17(3): 73-76. (2002)
4. Bourigault D. & Fabre C., Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de grammaire, Vol.25, pp.131-151. (2000)
5. Cao T-D. & Gandon F., Integrating external sources in a corporate semantic web managed by a multi-agent system. AMKM 2003 Stanford University. (2003)
6. Ciravegna F. & Petrelli D., User Involvement in Adaptive Information Extraction: Position Paper in Proc. of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining. (2001)
7. Corby O. & Faron-Zucker C., Corese: A Corporate Semantic Web Engine. WWW11 Workshop on Real World RDF and Semantic Web Applications, Hawaii. (2002)
8. Cunningham H. et al., GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. ACL'02. (2002)
9. Golebiowska J., Dieng-Kuntz R., Corby O. & Mousseau D., Building and Exploiting Ontologies for an Automobile Project Memory. K-CAP 01, Victoria, Canada. (2001)
10. Gruber T., A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition , p. 19-220. (1993)
11. Handschuh S., Koivunen M., Dieng R. & Staab S., eds, KCAP'2003 Workshop on Knowledge Markup and Semantic Annotation, Sanibel, Florida, October 26. (2003)
12. Kim S. et al., Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web. SAAKM'02, pp. 1-6, Lyon, France (2002).
13. Lassila O. & Swick R., W3C Resource Description Framework www.w3.org/TR/REC-rdf-syntax/ (2001)
14. Lindberg D., Humphreys B. & Mccray A. "The Unified Medical Language System", Methods Inf Med, pp 281-291 (1993)
15. McGuinness D.L. OWL Web Ontology Language Overview, www.w3.org/TR/owl-features/. (2004)
16. Nédellec C., Bibliographical Information Extraction in Genomics. IEEE Intelligent Systems & their Applications, 17(3):76-80, March/April. (2002)
17. Proux, D., et al. A Pragmatic Information Extraction Strategy for gathering Data on Genetic Interaction, in proceedings of ISMB (2000).
18. Schulze-Kremer S. & Smith B. & Kumar A., Revising the UMLS Semantic Network.(2002)
19. Séguéla P.,Aussenac-Gilles N., " Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine ". Actes de la conférence Ingénierie des Connaissances (IC'99). pp 79-88. Palaiseau. (1999)
20. Shatkay H., Edwards S. & Boguski M., Information Retrieval Meets Gene Analysis. IEEE Intelligent Systems & their Applications, 17(3):45-53. (2002)
21. Soo V., Lee C., Li C., Chen S. & Chen C., Automated Semantic Annotation and Retrieval Based on Sharable Ontology and Case-based Learning Techniques, JCDL'03, Texas. (2003)
22. Staab S., eds., Mining Information for Functional Genomics. IEEE Intelligent Systems & their Applications, 17(3):66-80, March-April. (2002)
23. Vargas-Vera M.et al. MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup, EKAW 2002, Siguenza, Spain (2002)
24. Weinstein J. N., et al., An information-intensive approach to the molecular pharmacology of cancer. Science 275:343-349. (1997)
25. Zweigenbaum P., MENELAS: an access system for medical records using natural language. Computing Methods Programs Biomed.45 (1-2):117-20. (1994)