Journal of Universal Computer Science, vol. 13, no. 12 (2007), 1881-1907
submitted: 15/4/07, accepted: 31/7/07, appeared: 1/12/07 © J.UCS

An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain

Khaled Khelif
(INRIA Sophia Antipolis, France
khaled.khelif@sophia.inria.fr)

Rose Dieng-Kuntz
(INRIA Sophia Antipolis, France
rose.dieng@sophia.inria.fr)

Pascal Barbry
(IPMC, Sophia Antipolis, France
barbry@ipmc.fr)

Abstract: This paper describes an ontology-based approach aiming at helping biologists to annotate their documents and at facilitating their information retrieval task. Our approach, based on semantic web technologies, relies on formalised ontologies, semantic annotations of scientific articles and knowledge extraction from texts. We propose a method/system for the generation of ontology-based semantic annotations (MeatAnnot) and a system allowing biologists to draw advanced inferences on these annotations (MeatSearch). This approach was proposed to support biologists working on DNA microarray experiments in the validation and the interpretation of their results, but it can probably be extended to other massive analyses of biological events (as provided by proteomics, metabolomics…).

Keywords: Semantic web, ontologies, semantic annotation, life science, NLP, Corese.
Category: H.3.1, H.3.3, M.0, M.7

1 Introduction

A considerable amount of knowledge is stored in textual documents (articles, research reports…) published on the web. This knowledge is essential for checking, validating and enriching new research work. But due to the large amount of data, from sources that are either internal or external to users' organisations, an efficient detection, storage and use of this knowledge is quite a major task. This is especially true for researchers manipulating huge amounts of biological data; for example in DNA microarray[1] experiments, several hundreds of experimental conditions can be analysed against 100,000 probes, and must be linked to thousands of scientific articles or reports. In such situations, biologists need tools supporting them for interpretation and/or validation of their experiments, which would ultimately facilitate planning of further experiments.

[1] http://www.gene-chips.com/

Our hypothesis is that this knowledge management problem can be solved by using semantic web technologies: ontologies can be used in order to provide formal grounding for representing the semantics of knowledge elements; they can guide creation of semantic annotations constituting a set of all meta-level characterisations easing knowledge source description, evaluation, and access.

The MEAT (Memory of Experiments for Analysis of Transcriptome) project [Khelif et al. 2005] [Khelif 2006] developed in collaboration with biologists working on the Nice Sophia Antipolis DNA Microarray platform (located at the IPMC[2] laboratory) aims at supporting biologists working on DNA microarrays. Its goal is to offer methodological and software support based on semantic web technologies (ontologies, semantic annotations) in order to ease interpretation and validation of DNA microarray experiments.

In this paper, we propose an ontology-based approach for generation of semantic annotations (the MeatAnnot system) and for information retrieval (the MeatSearch system).

1.1 Context

The DNA Microarray (or biochip) technology has been developed after the full sequencing of many genomes in order to get information about gene functions under many different biological contexts. Typical microarray experiments can assess thousands of genes simultaneously. Thus, they lead to a huge amount of information making it hard for a biologist to validate and interpret the obtained results.

For each biochip project, the involved biologists construct a textual corpus of papers concerning genes supposed a priori interesting for the microarray experiment carried out in this project. Of particular interest is the selection by the biologists of review articles, such as those provided by series or found on the web. Such a selection is useful, as it offers overviews of a specific field, overview written by a specialist of this field, and selected by another specialist (i.e. the biologist performing the microarray experiments). This corpus is then used in the validation/interpretation phase of experimental results.

The needs expressed by our partners biologists can be summarised as follows:

- *Support to validation of experimental results*: the biologist needs to search documents about the studied phenomenon so as to find information which argues, confirms or invalidates his/her assumptions; this implies the need of an accurate information retrieval system and requires rich annotations.
- *Support to interpretation of experimental results*: the biologist aims at identifying new/known relations or/and interactions between genes, cellular components and biological processes; this requires a view on knowledge contained in documents related to the experiment and inference capabilities over the annotations.

In the semantic web context, this information retrieval task can be carried out by associating to each document an RDF graph which gathers several annotations

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

extracted from its text and based on a domain ontology. These annotations describe the knowledge embedded in sentences containing ontology instances (concepts and relations) and constitute the link between entities in the textual documents and their semantic descriptions represented in the ontology. A semantic search engine using the ontology can draw rich inferences on these annotations; it allows to: (i) retrieve relevant documents, and (ii) reason on knowledge described by semantic annotations.

1.2 Motivations

As described below, the major need of biologists is to access knowledge described in natural language texts. Mining this literature is one way to detect relevant information and generate semantic annotations on documents in order to facilitate their search.

Our goal is to facilitate the information retrieval task for biologists. Therefore, in order to be able to create relevant annotations, we first asked: what information would a biologist be interested in, when reading an article. We thus studied how a biologist annotates a document: 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 their ways of annotating were different. The information selected by the different biologists was almost the same. They primarily underlined the names of the studied genes, substances or proteins, the studied biological phenomenon or the cellular functions as well as the verbs describing a relation between these various elements.

An example of sentence annotated by the three biologists was: *"KGF causes alveolar epithelial type II cell proliferation"*; this sentence asserts that the substance KGF causes a type of cell proliferation.

The representation of this kind of annotation must be well defined, easy to understand by all biologists and unambiguous. To fulfil these requirements, this annotation should be based on a formal model of the biomedical domain (e.g. an ontology).

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

In addition, the fact that these annotations are based on ontology incites us to use standard formalisms such as RDF(S) [McBride 04] or OWL [McGuinness 04] which allow the reuse of these annotations by different annotation tools and search engines.

The approach chosen in this work is to reuse existing ontologies in order to support a text mining method applied on biomedical literature. These ontologies define the type of entities and relations that we aim to discover through text analysis and they allow generation of rich annotations about documents. These annotations can then be used to perform information retrieval task.

Figure 1 shows the different stages of this approach.

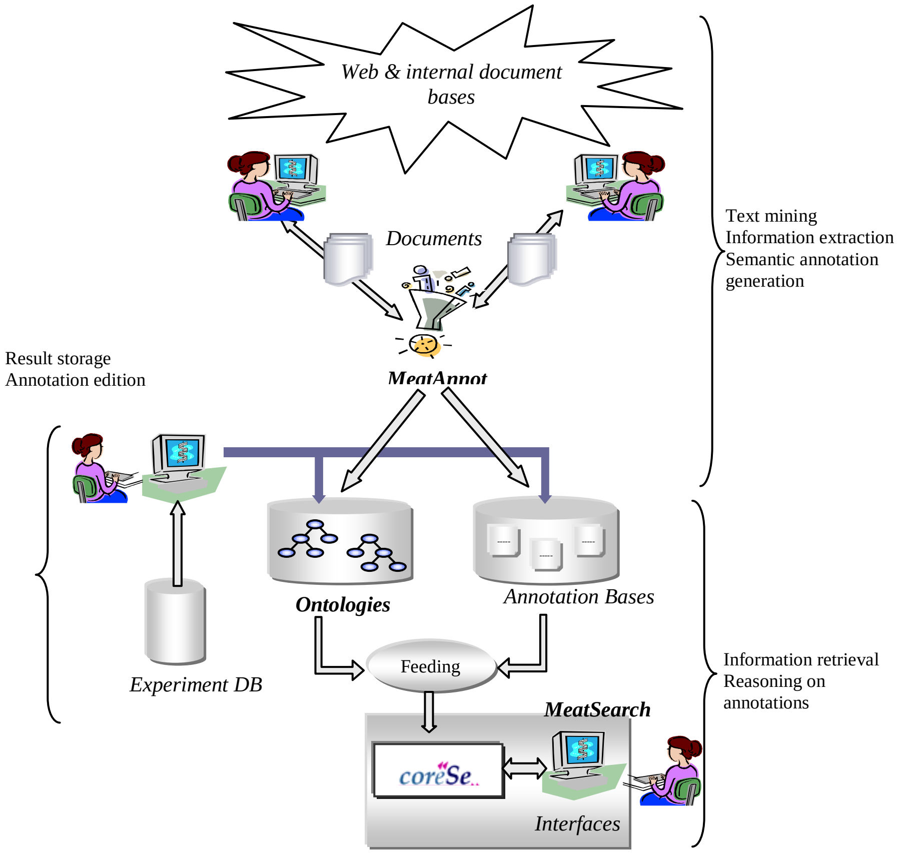


Figure 1: MEAT ontology-based approach for generating and using semantic annotations

2 An ontology for the biomedical domain

Like in most research domains, biologists aim to represent, share and reuse their knowledge. Therefore, several terminological systems were proposed and developed: controlled vocabularies for annotating genes [Ashburner et al. 2000] and indexing documents, thesauri[3] for navigating among domain terms and for easing information retrieval. As a step further, the biomedical community was interested in ontologies which aim at representing knowledge independently of any specific use[4]. Ontologies provide an organisational framework of the concepts and a system of hierarchical and associative relationships of the domain. In addition to the possibility of reuse and sharing allowed by ontologies, the formal structure coupled with the hierarchy of

[3] www.nlm.nih.gov/mesh/

[4] We must notice that this independence on the ontology w.r.t. the application is strongly criticised by researchers such as the French TIA working group.

concepts and the hierarchy of relations between concepts offers the possibility to draw complex inferences and reasoning.

2.1 The UMLS project

The Unified Medical Language System (UMLS) project was initiated in 1986 by the U.S. National Library of Medicine (NLM). Its goal is to help health professionals and researchers to use biomedical information from a variety of different sources [Humphreys and Lindberg 1993]. It consists of a (1) metathesaurus which collects millions of terms belonging to nomenclatures and terminologies defined in the biomedical domain and a (2) semantic network which consists of 135 semantic types and 54 relationships.

The semantic network represents a high-level abstraction for the metathesaurus; it is organised by distinguishing entities and events in two single-inheritance hierarchies. Each semantic type in the network has a textual definition and appears in one of these hierarchies.

The generation of ontology-based annotations on documents requires a lexicon of terms for referring to entities in the domain and an ontology describing this domain. For our case, this ontology must cover the entire biomedical domain (drugs, cells, genes, process…), but we noticed that except UMLS, all other ontologies were developed for a specific case (for example, GALEN [Rector et al. 1996]: clinical domain, MENELAS [Zweigenbaum 1994]: coronary diseases, GO [Ashburner et al. 2001]: molecular biology, etc.).

So, we chose the UMLS semantic network (SN) defined by [McCray 2003] as upper-level ontology for the biomedical domain: the hierarchy of semantic types can be regarded as a hierarchy of concepts and the terms of the metathesaurus as instances of these concepts.

In addition, GO has recently been integrated into UMLS [Lomax and McCray 2004]. Overall, a total of 23% of the GO terms either match directly (3%) or are linked (20%) to existing UMLS concepts. All GO terms now have a corresponding UMLS concept. This integration offers an important link between medicine and genomics terms.

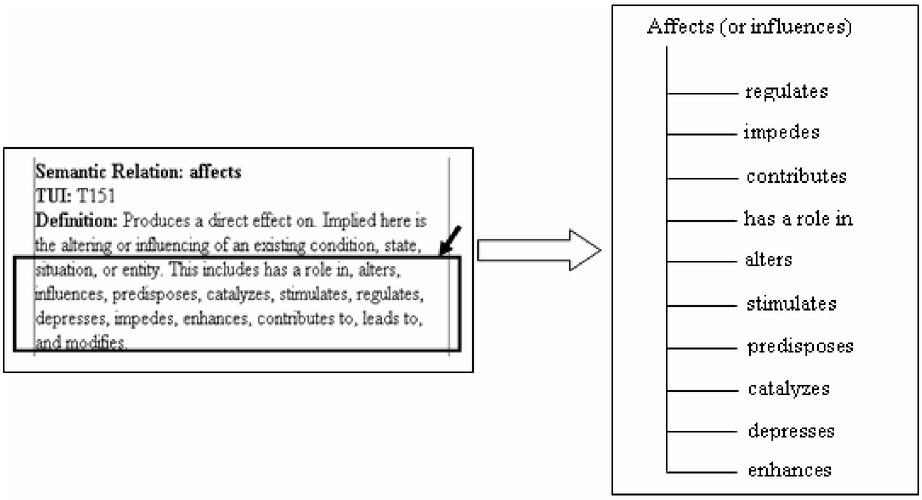
2.2 Enrichment of the relationships hierarchy

The UMLS semantic network comprises a hierarchy of 54 semantic relations made up of five families:

- Physical relations: connecting terms having common physic characteristic (example: branch_of);
- Space relations: connecting the terms according to their localisation (example: location_of);
- Functional relations: expressing a function or an activity connecting the terms (example: interacts_with);
- Temporal relations: connecting terms in time (example: precede);
- Conceptual relations: connecting terms according to some abstract concept, thought, or idea (example: measures).

After a thorough study of these various families and discussions with our colleagues biologists, it appeared that (i) to annotate a biological phenomenon, the two families primarily interesting for them are: conceptual relations and functional relations (65% of the whole set of the relations), and (ii) although these relations cover the totality of links that may exist between the concepts of the semantic network, some are too generic and can lead to a negative effect on the level of precision of an annotation.

For example, the functional relation 'affects' is defined by the production of a direct effect by a biological entity on another, this effect can be the result of the one of the following actions: {has a role in, alters, influences, predisposes, catalyzes, stimulates, regulates, removes, pressure from, impedes, enhances, contributes to}. This definition recommends that all these actions can be regarded as 'synonymous' and must be annotated by the relation 'affects'. But this can generate noise in the annotations using this relation: for example, a biologist aiming to find all the biological entities stimulated by a particular gene, will have in addition to the correct entities, others which were deteriorated, catalysed... by this same gene.



Semantic Relation: affects
TUI: T151
Definition: Produces a direct effect on. Implied here is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences, predisposes, catalyzes, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.

Figure 2: Example of relation enrichment: 'affects'

Our goal is to use this ontology to annotate resources and to facilitate the task of information retrieval; therefore we decided to enrich the semantic network by more specific relations in order to have more precise annotations. So, we proceeded in two steps:

1. Using relationships definitions

In this step, we relied on the definitions of each relation in the semantic network. As shown in figure 2, these definitions comprise a set of terms which can indicate a more precise sense of the concerned relation. These terms cannot be considered as synonyms of the concerned relation and some terms must be rather considered implicitly as more precise semantic relations. This assumption allowed us to specialise the UMLS relations by new relations. Figure 2 shows the result of this specialisation on the relation 'affects'.

2. Biologists' suggestions
During our discussions, the biologists proposed some relations specific to their field and which do not appear in the lists of terms characterising the UMLS relations. This step enabled us for example to add the relation 'activates' and the relation 'inhibits' as two specialisations of the relation 'performs'. We thus succeeded in adding 24 new relations to the semantic network of UMLS. These new relations have the same signature as the relation to which they are attached.

Finally, we developed a script which translates each semantic type and each relation from its textual format towards the corresponding concept and property in RDFS language. Two semantic types (respectively relations) linked by an *is-a* link in the UMLS SN are translated into two RDFS classes linked by subClassOf (respectively subPropertyOf) property. In addition, we used some primitives of OWL Lite (restriction, cardinality, etc.) to resolve some problems (discussed in [Khelif 06], [Kashyap and Borgida 2003]) occurring in the definition of the signature (domain and range) of relations.

3 UMLS-based semantic annotation generation: MeatAnnot

3.1 Method

In spite of its advantages, the creation of semantic annotations is a difficult and time-consuming process for biologists. Therefore, we developed a system called MeatAnnot which, starting from a textual document (i.e. a scientific paper), allows to generate a structured annotation, based on UMLS SN, and describes the semantic content of this text.

MeatAnnot uses the NLP (Natural Language Processing) tools GATE [Cunningham et al. 2002], TreeTagger [Schmid 1994], RASP [Briscoe and Caroll 2002] and our own extensions dedicated to extraction of semantic relations and of UMLS concepts. It processes texts and extracts interactions between genes and other UMLS concepts.

So, for each sentence, it tries to detect an instance of an UMLS relation and to detect the instances of UMLS concepts linked by this relationship and it generates an annotation describing this interaction (see more details in [Khelif 06]).

The generation method is decomposed in three steps described below:

Step1: Relation detection
In this step, we used JAPE [Cunningham et al. 2002], a language based on regular expressions and allowing us to write information extraction grammar for texts processed by GATE. So, for each UMLS relation (such as interacts_with, expressed_in, disrupts…), an extraction grammar was manually created to extract all instances of this relation.

The example below shows a grammar which allows detection of instances of the semantic relation "has_a_role_in" with its different lexical forms in the text (e.g. has a role, had roles, plays a positive role, etc.).

Example of grammar:

```
Rule:Has_role
 Priority: 1
(

 ({Tag.lemme == "have"}  |
  {Tag.lemme == "play"} )
  {SpaceToken}
 ({Tag.lemme == "a"}   |
  {Tag.lemme == "an"})
  {SpaceToken}
 ({Tag.cat == "JJ"}  {SpaceToken})?
  {Tag.lemme == "role"}
  {Tag.lemme=="in"})

  ):has_role -->
:has_role.RelationShip = {kind = "has_role", rule=Has_role}
```

In the above figure, Tag.lemme corresponds to the lemmatised form of the verb and Tag.cat corresponds to the grammatical category (JJ:adjective = important, vital, critical, etc.) of the term which can be present between the verb and the term 'role' ("?" means that it is optional).

Step2: Term extraction

To extract terms, MeatAnnot uses the Tokeniser module of GATE and the TreeTagger. The tokeniser splits text into tokens, such as numbers, punctuation and words, and the TreeTagger assigns a grammatical category (noun, verb...) to each token.

After tokenising and tagging texts, MeatAnnot uses an extraction window of four (four successive words are considered as a candidate term) and for each candidate term, if it exists in UMLS, MeatAnnot processes the following word, otherwise it decreases the size of the window till zero.

To interrogate UMLS, MeatAnnot uses the UMLSKS (the UMLS Knowledge Server based on the MetaMap[5] concept mapping program). This server provides access and navigation in the UMLS metathesaurus and in the UMLS semantic network. If the term exists in UMLS, the answer is obtained in XML format. This answer is parsed to obtain information about the term (semantic type, synonyms…); all this information is then used to generate the semantic annotation.

[5] http://mmtx.nlm.nih.gov/

In this step we noticed that MeatAnnot cannot detect some gene names because of the increasing number of gene synonyms. To solve this problem, the biologists supplied us with a dictionary of specific genes used frequently in DNA experiments. So, after the extraction phase, MeatAnnot re-processes the text and tries to detect missing genes.

Some other specific biomedical terms were not detected by MeatAnnot (not found in UMLS).

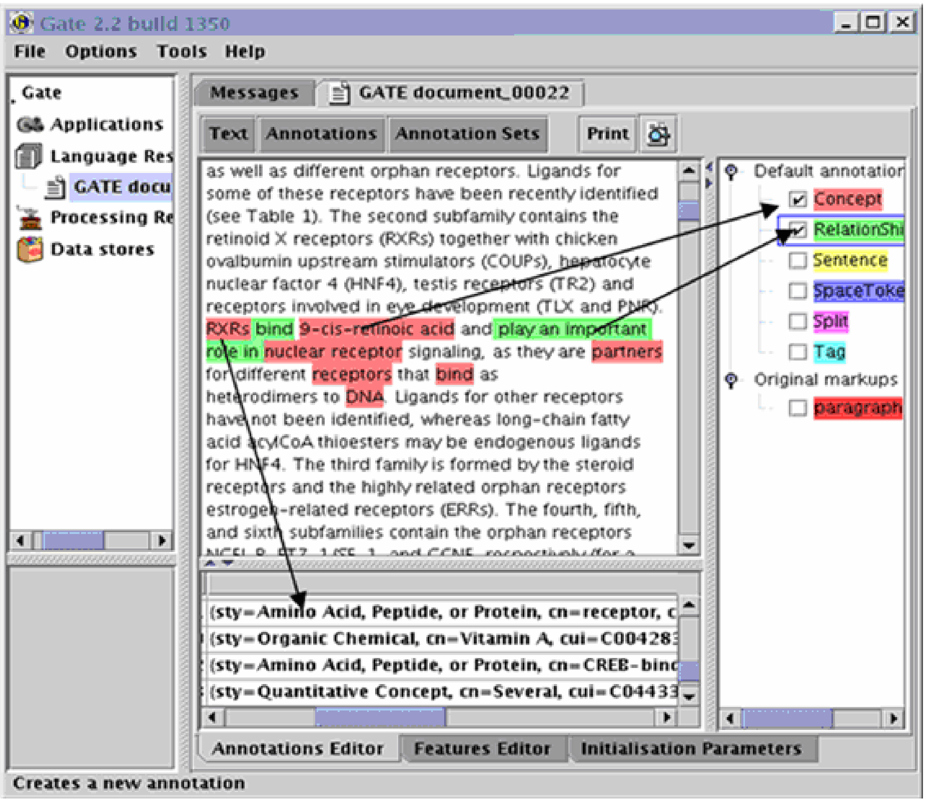
Example of sentence: "*ERK-5 also plays a role in the AP-1 regulation*"

In this sentence, MeatAnnot generates an annotation describing the relation (has_role_in) between the two genes *ERK-5* and *AP-1* since it cannot detect the term *AP-1 regulation* (if it does not exist in UMLS); this annotation is wrong or not relevant for biologists. Therefore, we developed some heuristics to solve this kind of problems:

```
H1:     {term1.sty == 'Gene_or_Genome'}
        {term2.string ∈ GF_termes} =>
        {term3 = term1+term2; term3.sty = 'Genetic_Function'}

GF_termes ={'induction','translation','regulation','expression','mutation','deletion'}

H2:     {term1.sty == 'Amino_acid_Peptide_or_Protein'}
        {term2.string ∈ MF_termes } =>
        {term3 = term1+term2; term3.sty = 'Molecular_Function'}

MF_termes ={'activity','binding','phosphorylation'}
….
```

H1 implies that, if a term detected as a gene instance is followed for example by the word "*regulation*", we can consider that the concatenation of both words is a 'Genetic_function' instance.

These heuristics can help to improve the term extraction phase and to enrich the UMLS metathesaurus with new terms and their associations to the SN.



as well as different orphan receptors. Ligands for some of these receptors have been recently identified (see Table 1). The second subfamily contains the retinoid X receptors (RXRs) together with chicken ovalbumin upstream stimulators (COUPs), hepatocyte nuclear factor 4 (HNF4), testis receptors (TR2) and receptors involved in eye development (TLX and PNR). RXRs bind 9-cis-retinoic acid and play an important role in nuclear receptor signaling, as they are partners for different receptors that bind as heterodimers to DNA. Ligands for other receptors have not been identified, whereas long-chain fatty acid acylCoA thioesters may be endogenous ligands for HNF4. The third family is formed by the steroid receptors and the highly related orphan receptors estrogen-related receptors (ERRs). The fourth, fifth, and sixth subfamilies contain the orphan receptors

```
(sty=Amino Acid, Peptide, or Protein, cn=receptor, c
(sty=Organic Chemical, cn=Vitamin A, cui=C004283
(sty=Amino Acid, Peptide, or Protein, cn=CREB-binc
(sty=Quantitative Concept, cn=Several, cui=C04433
```

Figure 3: Example of relation detection and term extraction

Figure 3 is a GATE interface showing the obtained result after the two steps of relation detection and term extraction. In this example, two relations (bind and play_role) and seven terms were detected in this sentence.

Step3: Annotation generation

In this step, MeatAnnot uses the RASP module which assigns a linguistic role (grammatical relation) to sentence words (subj, obj …): it allows finding out concept instances linked by the relation.

So for each detected relation, MeatAnnot analyses the results of the extraction phase and checks if the subjects and objects of this relation were detected as UMLS concepts. Then it generates an annotation describing an instance of this relation.

Since RASP processes only single words, the linguistic roles of multi-terms are deduced automatically by MeatAnnot. For example, in the sentence *”KGF causes lung injury”*, RASP first assigns the object role to *”injury”* but MeatAnnot re-assigns this role to *”lung injury”* since it detected it as a UMLS concept instance.

The example below summarises the process steps. Let us consider the sentence: *“IFN-alpha and IFN-beta are secreted by dendritic cells.”*

First: by applying the extraction grammars on this sentence, MeatAnnot detects (by the presence of verb *'to secrete'*) that it contains the UMLS relation "produce".

Second: Table 1 describes the result of the term extraction phase.

Term	Semantic type	Synonyms
IFN-alpha	Amino Acid Peptide or Protein	alfa-n3 interferon, Ginterferon, G-interferon, et.
IFN-beta	Amino Acid Peptide or Protein	Endogenous Interferon Beta, IFNb, IFN-B, etc.
dendritic cells	Cell	N/C

Table 1: Term extraction results

Third: MeatAnnot applies the RASP module on the sentence and parses the result to detect the different linguistic roles of the words.

An excerpt of the result of RASP on this sentence is:

```
(|ncsubj| |secrete+ed:5_VVN| |IFN-alpha:1_NN1| |obj|)
(|ncsubj| |secrete+ed:5_VVN| |IFN-beta:3_NN1| |obj|)
(|arg_mod| |by:6_II| |secrete+ed:5_VVN| |cell.+s:8_NN2| |subj|)
(|conj| _ |IFN-alpha:1_NN1| |IFN-beta:3_NN1|)
(|ncmod| _ |cell.+s:8_NN2| |dendritic:7_JJ|)
(|aux| _ |secrete+ed:5_VVN| |be+:4_VBR|)
```

Lines 1, 2 and 3 indicate that (i) the words "*IFN-alpha*", "*IFN-beta*" and "*cells*" are linked by the verb *"secrete"*, and (ii) the linguistic role affected to IFN-alpha and *IFN-beta* (resp. *cells*) is *object* (resp. *subject*).

"*dendritic cells produce IFN-alpha*" and "*dendritic cells produce IFN-beta*" are thus detected as instances of the relation produce; so, MeatAnnot generates an RDF annotation for these two instances and adds it to the annotation concerning this paper.

```
<m:Cell rdf:about='#dendritic_cells'>
        <m:produce >
                <m:Amino_Acid__Peptide__or_Protein rdf:about='#IFN-alpha'/>
        </m:produce>
        <m:produce >
                <m:Amino_Acid__Peptide__or_Protein rdf:about='#IFN-beta'/>
        </m:produce >
</m:Cell >
```

After text processing, MeatAnnot generates an RDF annotation describing all these interactions described in the article and stores it in the directory containing the annotations of the other papers. Each article is linked to the RDF file containing its annotations. The current system has a flat annotation base; this base can be organised in the future, for example according to the article theme or to the user supplying with the corpus.

These annotations can then be used, either in a bibliographical search or in a more complex IR (Information Retrieval) scenario such as searching interactions between genes or of genes with other biomedical entities.

3.2 Evaluation

To validate our annotations, we adopted a user-centred approach: we chose randomly a test corpus (2751 sentences) from the documents given by biologists and we presented the suggestions proposed by MeatAnnot to biologists via an interface in order to evaluate their quality.

Since these annotations were intended for an IR context, we focused on classic IR quality measures for indexing and we adapted them to our case.

We noticed also that some suggestions were considered as correct but not useful to the biologists since they described a basic or vague knowledge. Therefore, we introduced a new measure, called *usefulness,* for measuring the rate of useful suggestions. This measure is subjective because it relates to a point of view of a user or of a group. In this work, the annotations considered as useless by a biologist are stored in the annotations base. A possible improvement would be to add metadata on these annotations (for example: *useless_for*) which would allow to filter the answers sent by this biologist.

	Measures
Precision	$\frac{\text{Nb suggestions correctly extracted}}{\text{Nb all suggestions extracted}}$
Recall	$\frac{\text{Nb suggestions correctly extracted}}{\text{Nb suggestions that should be extracted}}$
Usefulness	$\frac{\text{Nb useful suggestions extracted}}{\text{Nb suggestions correctly extracted}}$

Table 2: Measures for the quality of the annotations

Precision relates to the absence of noise (also called commission) in the extraction and *recall* relates to the absence of silence (also called omission).

	Suggestion	Correct	Missing	Useful	Precision	Recall	Usefulness
Result	509	426	274	399	0.836	0.608	0.936

Table 3: Quality of Meatannot suggestions

The second column describes the number of relations correctly extracted from texts. The difference with the number of suggestions proposed by MeatAnnot is mainly due to the errors generated by the NLP tools (e.g. wrong grammatical category or wrong linguistic role) and to the terms missing in UMLS (i.e. when the subject or object of a relation was not found in UMLS). Nonetheless a good precision is obtained since 83% of the suggestions were correct.

The third column describes the number of relations not extracted by MeatAnnot: these missing suggestions are also due to the errors generated by the NLP tools and mainly to relations deduced by the biologist (when s/he reads the sentence) and which cannot be generated automatically.

Example of errors generated by the NLP tools:

"TRP gene, which belongs to the TRP-homolog group, is expressed in neurons"

In this sentence where the relation "expressed_in" is detected, the RASP module suggests that *"which"* is the subject of the relation, so MeatAnnot does not generate the annotation because *"which"* is not an UMLS term and it loses the interaction between the *"TRP gene"* and *"neurons"*.

Example of missing relations:
"Upon interferon-gamma induction, after viral infection for example, a regulator of the proteasome, PA28 plays a role in antigen processing."

In this example, MeatAnnot extracts automatically the relation "PA28 plays_role antigen processing" but a biologist who reads this sentence can deduce, using his/her implicit knowledge, another relation which is "interferon-gamma have_effect PA28".

Finally, MeatAnnot has a good usefulness since 93% of correct suggestions are considered as useful by biologists. The annotations regarded as useless are however added to the RDF file containing the other annotations: they have no negative impact and they may be relevant to novice or non expert users.

These results prove that MeatAnnot generates good quality annotations, an essential feature for a use in an information retrieval context.

For the real-world application, we applied our technique on the Generif[6] corpus (about 11540 documents) which provides texts describing human genes using concise phrase.

3.3 Towards a generic methodology

We presented a method based on semantic web technologies for generation of ontology-based semantic annotations for biological domain. This method can be generalised to any other scientific domains (chemical domain, physical domain, etc.); since they have the same needs such as the support for the automatic generation of

[6] http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html

rich annotations from texts, which can help to validate and to interpret experimental results.

In fact, the modules presented are reusable and rely on standard technologies. The MeatAnnot method can be generalised in four points:

- Selection of an ontology covering the domain studied;
- Development of an API to query the ontology. MeatAnnot proposes a module which can detect concepts labels of any ontology; it relies on Corese and on our term extraction algorithm.
- Definition of the extraction grammar for the ontology relations; It needs a small study about the definitions of relations in the ontology and about its linguistic forms in texts.
- Reuse of MeatAnnot modules to detect terms and relationships and then generate semantic annotations. It requires to define the chosen annotation schema.

Remarks:

(1) If the selected ontology needs enrichment or population by instances, it is possible to enrich it by using NLP tools on a textual corpus provided by the domain experts.

For example:

- Using Nomino[7] or Likes[8], to enrich and populate the concept hierarchy (as in the SAMOVAR system [Golebiowska et al. 2001].
- Using Syntex [Bourigault and Fabre 2000], to extract verb syntagms considered as relevant for the domain and which enable to enrich the relation hierarchy.

(2) The development time for adapting the application depends on the number of the relations in the ontology and on the chosen annotation schema.

Figure 4 recapitulates the obtained method.

[7] http://www.ling.uqam.ca/nomino

[8] http://www-ensais.u-strasbg.fr/liia/likes/likes.htm

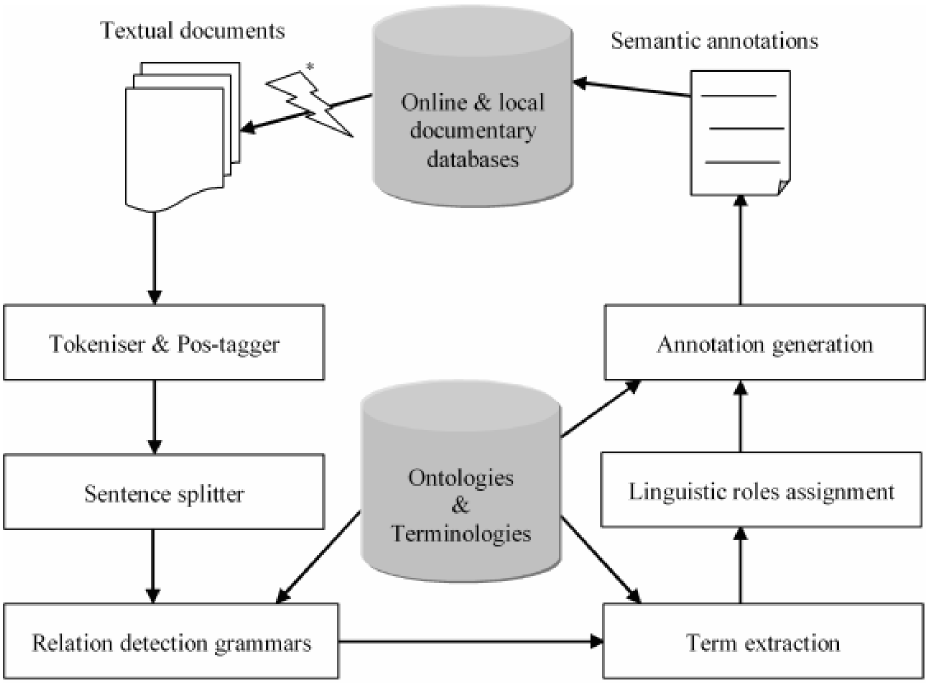

Textual documents
*
Online & local documentary databases
Semantic annotations
Tokeniser & Pos-tagger
Annotation generation
Ontologies & Terminologies
Sentence splitter
Linguistic roles assignment
Relation detection grammars
Term extraction

Figure 4: Ontology-based semantic annotation generation method

**: conversion of documents from their original format (generally PDF) to textual format.*

This method allows the generation of semantic annotations based not only on concept instances but also on relation instances. In addition to document description, these annotations embed information about domain knowledge.

4 Annotation-guided search: MeatSearch

For enabling the biologists to use these annotations, we developed a tool called MeatSearch based on the semantic search engine CORESE (Conceptual Resource Search Engine) [Corby et al. 2004], [Corby et al. 2006] and composed of a set of GUI allowing users to ask queries on the annotation base (see Figure 5).

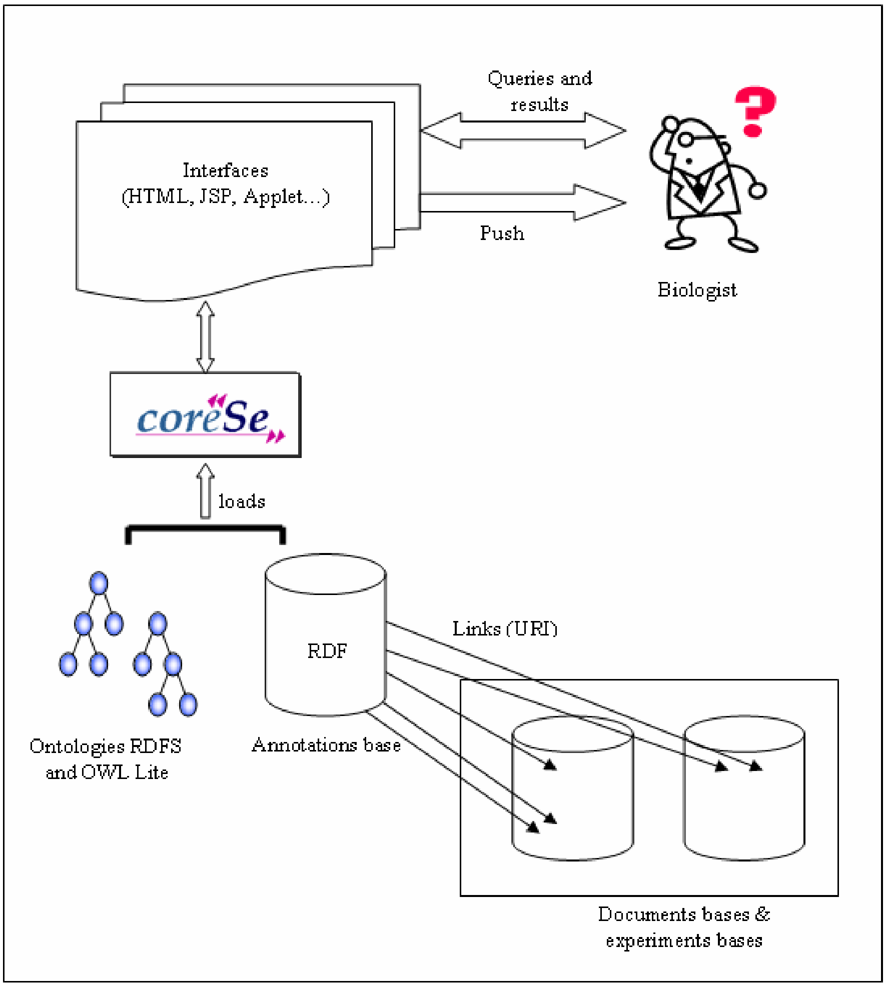

Queries and results
Interfaces (HTML, JSP, Applet...)
Push
Biologist
coreSe
loads
RDF
Links (URI)
Ontologies RDFS and OWL Lite
Annotations base
Documents bases & experiments bases

Figure 5: The MeatSearch architecture

Using adequate XSLT [Clark 1999] style sheets, MeatSearch transforms the CORESE results into graphical or/and textual presentation understandable by biologists. It also provides links to the sentence from which the annotation was extracted and to the document containing this sentence. This offers an interesting documentation on the annotations and this ability to trace the provenance is very useful for validation.

4.1 Use of CORESE

To formalise our ontologies and annotations, we chose RDFS [McBride 2004] and RDF [Lassila and Swick 2001] languages: they are recommended by W3C, respectively to represent light ontologies and to describe web resources using ontology-based semantic annotations.

This choice enabled us to use the semantic search engine CORESE which allows to:

- Navigate in the annotation bases taking into account the concept hierarchy and the relation hierarchy defined in the ontology.
- Add rules which complete the annotation bases.
- Reason on the whole annotation base constructed from different heterogeneous sources (papers, experiment database): the biologist can thus deduce implicit and explicit knowledge about a gene.
- Use different levels of access (admin, public, group…) to the annotation base.
- Have different views on the annotations.

The use of standards offers the portability of data and allows to rely on other semantic engines. [Hoang and Tjoa 2006] presents a state of the art of ontology-based query systems which can be used to implement tools like MeatSearch.

4.2 Examples of use

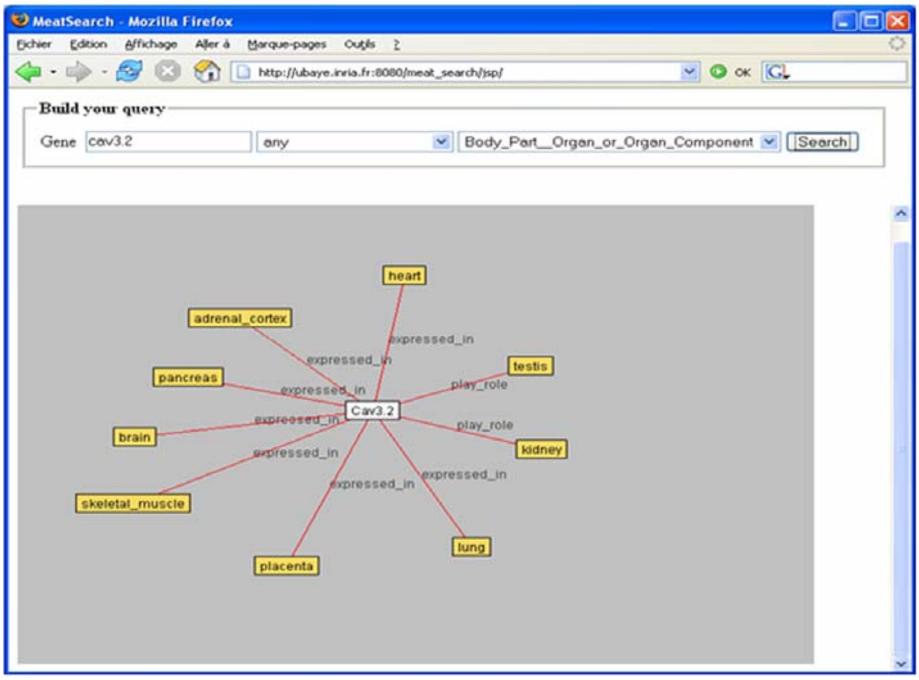
CORESE interprets SPARQL[9] queries (currently under discussion as a W3C candidate recommendation); it enables to write queries constituted of a boolean combination of RDF triples.

For example, the following query enables to retrieve all relations between a gene called "*cav3.2*" and a part of the human body:

```
select ?g ?r ?b where
{?g   rdf:type   m:Gene_or_Genome.}
{?g   =   'cav3.2'.}
{?g   ?r   ?b.}
{?b   rdf:type   m:Body_Part__Organ_or_Organ_Component}
```

This query is generated automatically by MeatSearch and the result is formatted in a graphical representation (see Figure 6 and Figure 7) to facilitate its visualisation.

[9] http://www.w3.org/TR/rdf-sparql-query/


MeatSearch - Mozilla Firefox
Fichier Edition Affichage Aller à Marque-pages Outils ?
http://ubaye.inria.fr:8080/meat_search/jsp/
OK

Build your query

Gene cav3.2 any Body_Part__Organ_or_Organ_Component Search

Figure 6: Previous query result presentation in MeatSearch

The MeatSearch interfaces allow biologists to build complex queries using simple graphical features and generate the adequate SPARQL query; they also provide links towards the documents described by the annotations (for example, in figure 6, each edge is linked to the sentence from which the interaction was extracted). User can navigate, for example, in this graph by clicking on nodes (MeatSearch generates a query describing all interactions with this entity) and on edges (MeatSearch shows the sentence from which the annotation was generated).

In addition, the biologists can be interested in any query aiming to find interactions between biomedical entities: genes and genes, genes and diseases, genes and proteins, etc. To do this, MeatSearch proposes a ‘free query’ interface which enables to generate formal queries and ask CORESE (see Figure 7).

CORESE offers a rule language [Corby et al. 2006] which enables to deduce new knowledge from existing one. The production rules are applied on the annotation base to complete it and to add more information in order to reduce silence in the IR phase.

Through discussions with our partners’ biologists, we produced such rules dedicated to DNA experiment memory.

An example of rule is:

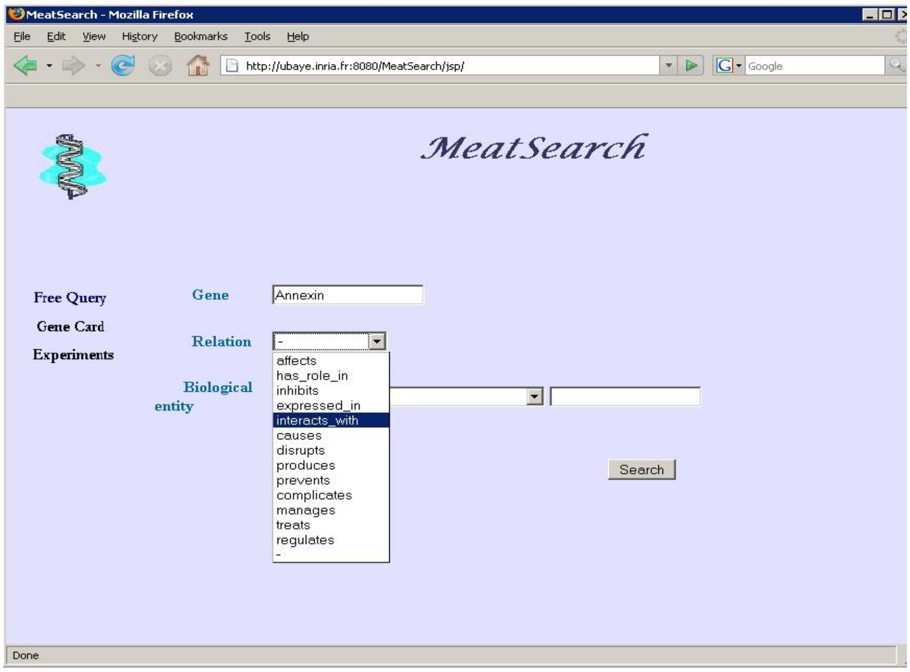
“For each receptor which activates a molecular function, if this function plays a role in an organism function, the receptor can play the same role”

This rule is expressed as:

```
IF    ?r    rdf:type    m:Receptor
      ?r    m:activates    ?mf
      ?mf   rdf:type    m:Molecular_Function
      ?mf   m:plays_role ?of
      ?of   rdf:type    m:Organism_Function
THEN
      ?r    m:plays_role    ?of
```

These rules enrich the annotation base and can improve the recall/precision of the information retrieval system.

The formalisation of such rules requires some knowledge in SPARQL/RDF modelling. So, to help biologists to enrich the rules base, we can imagine, in the future, (i) an ergonomic interface which guides the formalisation, or (ii) an intermediary language (semi-formal), easy to use by biologists and which can be wrapped automatically to SPARQL.


MeatSearch - Mozilla Firefox

http://ubaye.inria.fr:8080/MeatSearch/jsp/

Gene: Annexin

Relation: -

Figure 7: The free query interface of MeatSearch

A last example concerns the addition of metadata on an annotation so as to give more information on:

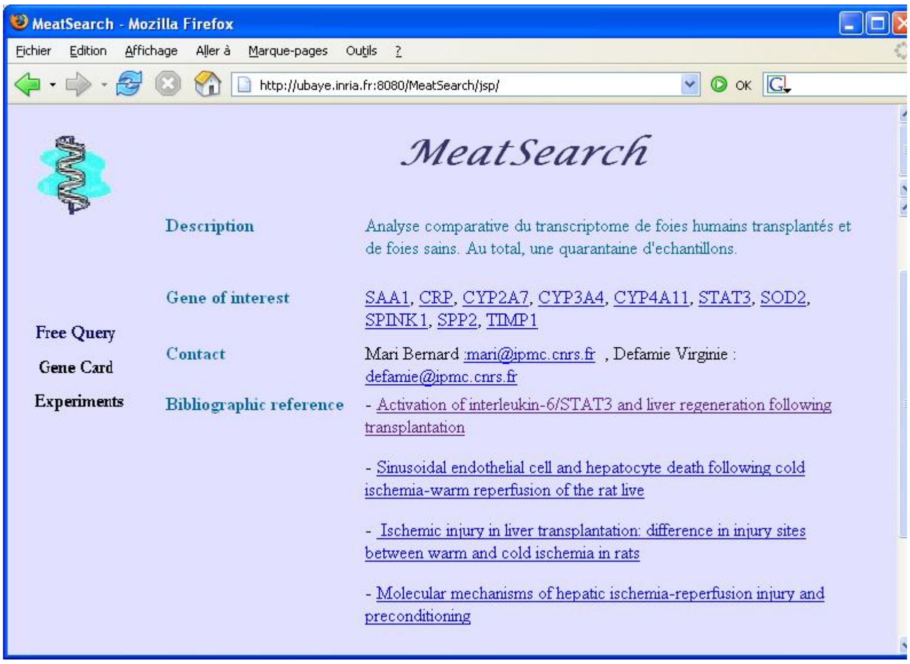
- The *source of the resource*: the biologist who supplied us with the paper to annotate or the biologist who performed the experiment.

- The *source of the annotation*: generated automatically by MeatAnnot vs. added/validated by a biologist.
- The *general topic of the annotation*: the different biologists may have different centres of interest about the same experiment.

The annotation below describes a paper given by a biologist named *'Bernard Marie'* and related to experiments on *'liver'*.

```
<do:paper rdf:about='http://www-sop.inria.fr/acacia/meat/livertransplantation.pdf'>
    <do:providedBy>Bernard Marie</do:providedBy >
    <do:describeExp rdf:resource='#experiments1934'/ >
    <do:relatedTo >
          <m:Body_Part__Organ_or_Organ_Component  rdf:about='#liver'/>
    </do:relatedTo >
    ....Annotation…
    <do:generatedBy>MeatAnnot</do:generatedBy>
    <do:validatedBy>Bernard Marie</do:validatedBy>
    …Annotation…
</do:paper>
```

Figure 8 shows the result of a query trying to find experiments (only papers describing experiments) related to *'liver'* and validated by a person named *'Bernard Marie'*.



Description: Analyse comparative du transcriptome de foies humains transplantés et de foies sains. Au total, une quarantaine d'echantillons.

Gene of interest: SAA1, CRP, CYP2A7, CYP3A4, CYP4A11, STAT3, SOD2, SPINK1, SPP2, TIMP1

Contact: Mari Bernard :mari@ipmc.cnrs.fr , Defamie Virginie : defamie@ipmc.cnrs.fr

Bibliographic reference:
- Activation of interleukin-6/STAT3 and liver regeneration following transplantation
- Sinusoidal endothelial cell and hepatocyte death following cold ischemia-warm reperfusion of the rat live
- Ischemic injury in liver transplantation: difference in injury sites between warm and cold ischemia in rats
- Molecular mechanisms of hepatic ischemia-reperfusion injury and preconditioning

Figure 8: Metadata query result presentation in MeatSearch

MeatSearch uses these metadata to propose different views on the annotation base related to users (annotation source), the context (general topic of the annotation) and the method of annotation generation (automatically vs. manually).

Queries on these metadata are very useful for browsing the annotation base, for checking its coherence and for the whole validation phase.

5 Conclusions

5.1 Discussion

In this paper, we presented an approach based on semantic web technologies for generation and use of ontology-based semantic annotations. The generated semantic annotations can be used in several scenarios, such as:

- Improvement of the document retrieval phase: the use of the concepts/relations hierarchies to expand users' queries improves recall.
- Discovering new knowledge: CORESE can find out paths between two entities. A path is constituted by a set of relations. In our case, biologists can deduce the role played by a selected gene in a disease by analysing the path found out between them (such a technique was used in semantic web service aggregation [Gandon et al. 2005]).

Another originality of our work consists of the use of several technologies to provide a real world Corporate Semantic Web Application that (i) relies on formal semantics (Ontologies, Semantic annotations) which reduce ambiguity compared to informal semantics, (ii) offers drawing inferences on these semantics at runtime (by using CORESE), and (iii) uses text to extract information (NLP tools) which is a very rich source of knowledge.

Last, we think that an evaluation study on the generated annotations (as the evaluation proposed in this paper) is necessary since this generation phase is expensive and often irreversible.

This paper proposes some solutions to problems raised during the final discussion of W3C Workshop on Semantic Web for Life Sciences [W3C 2004]:

- Good quality of the annotations extracted automatically: MeatAnnot annotations.
- Adequate representation of the context: our metadata on annotations which gives new ways of reasoning and more information on the annotation base.
- Possibility of reasoning on annotations: CORESE enables such reasoning.
- Semantic web browsing: we offer an automatic association of semantics to the knowledge resources and we provide a user-interface support.

Finally, the MeatAnnot module is used in several applications. For example, for annotating textual fields of patents [Ghoula et al. 2007] or for annotating a website pages in order to facilitate users profile detection [Mrabet et al. 2007].

5.2 Related work

The couple MeatAnnot/MeatSearch generates and uses ontology-based semantic annotations extracted from texts: these capabilities are close to those proposed by the Textpresso system [Muller et al. 2004]. Textpresso is an ontology-based information retrieval and extraction system for biological literature. It identifies terms (instances of the ontology concepts) by matching them against regular expressions and encloses them with xml tags. Textpresso also offers user interfaces to query these annotations. But Textpresso has the following drawbacks: (i) the annotation is embedded in the text, which makes difficult its reuse by other systems (while MeatAnnot generates an RDF annotation separate from the document), and (ii) it needs thousands of regular expressions to extract relevant terms (while in MeatAnnot, linguistic analysis performs the text matching task).

MeatAnnot has some similarities with other available text mining solutions which do not rely on ontologies. For example with GeneWays [Rzhetsky et al. 2004] which allows to select scientific journals, identify gene/protein names in the journal text, extract interactions between these gene/protein and other actions by means of NLP and storing these interactions in a database.

MeatAnnot uses the GATE API to process texts and uses a pre-populated ontology to extract terms. So, it can be compared to (i) the KIM system proposed by [Popov et al. 2004]. However, the semantic annotations generated by KIM are not used to annotate documents but to enrich an existing knowledge base, (ii) the PASTA system [Gaizauskas et al. 2003] which enables to extract information on the roles of amino acid residues in protein active sites.

Relying on NLP techniques, MeatAnnot differs from semantic annotation systems that use a machine-learning based information extraction approach. These systems (for example S-CREAM [Handschuh et al. 2002] and MnM [Vargas-Vera et al. 2002]) rely on manually annotated documents on which the method can be trained. For an overview on annotation systems, see [Uren et al., 2006].

Relation extraction was studied by [Séguéla and Aussenac-Gilles 1999] that propose the CAMELEON method/system which allows the extraction of semantic relations between terms using linguistic patterns (For example "*X is Indefinite_Article Y*" for hypernomy relation). This method relies on morpho-syntactic regularity in texts and needs a pre-processing phase to define specific patterns for a domain.

Our method also uses patterns (JAPE grammar) to detect relations but it relies on an advanced syntactic analysis of texts (cf. use of linguistic roles) to extract terms linked by the relation. Methods like CAMELEON could be used in our system as complement to improve the relation extraction phase (for example when the system fails to assign the correct linguistic role in a sentence).

To discover interactions among genes, [Nédellec 2002] proposes to use training corpora in order to generate extraction rules or patterns; these patterns are used in the extraction phase for annotation generation. This approach differs from our method since it is based not only on NLP tools but also on machine learning techniques.

An original theme-finding method is presented in [Shatkay et al. 2002]; it consists of characterising each theme by a set of term probability distributions. The algorithm then extracts terms from document abstracts and uses the distributions to classify them by theme; for example, documents discussing about genes responsible for nutrition in yeast, are likely to contain terms such as "*fructose*" or "*glucose*" and unlikely to contain the term "*lipid*".

An overview of different mining methods in the biomedical domain is presented in [Shatkay and Feldman 2003] [Staab 2002].

Finally, MeatSearch can also be compared with web reasoning systems [Ohlbach and Schaffert 2004] applied on corporate memory, since it integrates CORESE and enables advanced information retrieval and reasoning on annotations.

5.3 Lessons Learnt and Further Work

We can distinguish several kinds of lessons learnt: from conceptual and methodological viewpoint, from technical viewpoint and from applicative viewpoint.

5.3.1 Technical lessons learnt

In this work, we tested and used several NLP tools for building our information extraction system. The first problem raised in this phase was the component integration; this problem is due to the difference between the input/output formats of the different tools. We solved this problem by using the GATE API which (i) provides tools such as tokeniser, pos-tagger, gazetteer… and (ii) offers the possibility to integrate any new component (existing tools such as RASP, TreeTagger or our own extensions). Moreover, the conversion of articles from PDF format towards plain text is very problematic for several reasons: (1) PDFs are generated by several different tools, (2) biologists often use Greek alphabet characters that are difficult to recognise, (3) journals and books have different layouts for article presentation. For this phase conversion, we used an OCR (Optical Character Recognition) software; it gives good results but it requires user's intervention. So, it is necessary to develop an automatic converter taking into account all these problems.

5.3.2 Performance evaluation

We adopted a full text analysis approach while most of existing systems process only article abstracts. Our approach is clearly time-consuming (the complete processing of a page takes about 3 minutes) but it is worthwhile since it increases the recall/precision of the information extraction phase and gives more information about knowledge embedded in texts. The installation of a local version of UMLS and some technical optimisations can decrease the running time of the system. Moreover, as this phase is a batch preprocessing independent of the later real-time processing of any user query, it prepares more efficient query processing.

5.3.3 Discussion on ontology reuse

Our approach based on UMLS confirms that reusing existing domain ontologies can help to build real-world semantic web applications. In fact, despite some knowledge engineering problems (discussed above), the use of an existing upper-level ontology (such as UMLS SN) coupled with a rich terminology (UMLS metathesaurus) facilitates the information extraction process and allows to generate rich and shareable annotations. We think that the UMLS-based approach should be generalised to different domains needing interpretation and reasoning.

Several researchers have emitted doubts about possible reuse of ontology. They insist on the influence of the intended application on the ontology: some modelling choices or some ontology structuring choices are influenced by the future application aimed. But our experiment in MEAT project clearly showed the interest of using UMLS as reference ontology. In our work, the ontology was altogether the reference w.r.t. to which the annotations were created, the terms extracted and the relations extracted. To confirm the interest of such reference ontologies, for relations, we had first relied on Syntex tool that offered both term extraction and verbal syntagms extraction from a corpus of sample articles in biology. Our objective was to propose an extraction grammar for each relation indirectly expressed by these verbal syntagms (independently of any reference ontology, so as to offer a bottom-up approach, and to rely only on relations attested by texts). But it appeared that all the interesting relations were already included in UMLS relations. It confirms that a library of relation extraction grammars (written in JAPE) can be reusable by other researchers aiming at extracting UMLS relations from biomedical texts. The validation phase can be useful for indicating that some relations were not adequately extracted because of the lack of accuracy of the grammar for extracting this relation. So this phase can enable to refine this relation extraction grammar. A tool such as Syntex can be useful in this purpose.

The good results obtained in the information extraction phase confirm that the automation of this task is useful and it eases the user's work. In addition, the use of standard semantic web technologies for formalisation of this information into ontology-based semantic annotation can solve the knowledge sharing problem.

Even more, we think that this method can be adopted to annotate online documentary databases (such as Pubmed); the annotation base obtained might represent a very rich knowledge source for biologists. Nevertheless, we must not forget that an assumption underlying annotation generation by MeatAnnot is the consistency of the different articles and the absence of contradictions among them. This hypothesis enables to reason about the global RDF annotation base containing all the annotations stemming from the different articles, as in a global knowledge base. Therefore if we tackle the whole Pubmed articles, we must be vigilant about the coherence of the obtained annotations since contradictory biologist's viewpoints or wrong results may be contained in these articles.

5.3.4 Discussion on W3C standards

By generating RDF annotations, we rely on W3C recommendations for semantic Web. For queries and rules, there is not yet any official recommendation. However, a W3C working group works on SPARQL as future query language recommendation. CORESE query language – that we used in MEAT - is very close to SPARQL and handles most SPARQL features. Moreover, CORESE has the advantage to already offer processing of queries expressed in its language. Moreover it must be noticed that the MEAT end-users – biologists – use user interfaces for expressing their queries and do not directly handle this query language: these internal queries are generated automatically from the user interfaces. Concerning the rules, so far, there is not yet any rule language recommended by W3C. Therefore the best solution was to use CORESE rule language for which CORESE offers a rule engine that has been working quite efficiently since several years [Corby et al. 2004].

5.3.5 Further work

In future versions of MeatAnnot, we aim to take into account contextual information during the knowledge extraction phase. Let us take the example of the following sentence: *"In vitro assays demonstrated that only p38alpha and p38beta are inhibited by csaids"*. Actually, in this sentence, MeatAnnot identifies that *'p38alpha'* and *'p38beta'* are inhibited by *'csaids'* but does not detect the fact that this inhibition is observed *'in vitro'* whereas this information can be very important for the interpretation of a particular result.

MeatSearch can also be improved by introducing some typical search scenarios proposed by biologists.

Finally, like for most semantic web applications, we must propose solution to manage the ontology evolution. In fact, such evolution can cause inconsistencies in the annotation base, which induce errors in the information retrieval phase.

Acknowledgements

We thank Remy Bars from Bayers Cropscience, the IPMC team working on microarray experiments, especially Kevin Le Brigand for fruitful discussions, PACA[10] region which funded this work by a regional grant and Sealife IST project where this work is extended.

References

[Ashburner et al. 2000] Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet, 25, 25-29, (2000).

[Bourigault and Fabre 2000] Bourigault D. and Fabre C., Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de grammaire, Vol.25, 131-151, (2000).

[Briscoe and Caroll 2002] Briscoe E. and Carroll J., Robust accurate statistical annotation of general text. In Proceedings of the Third IC LR E, Las Palmas, Gran Canaria. 1499-1504 (2002).

[Clark 1999] Clark J., XSL Transformations (XSLT) Version 1.0, W3C Recommendation, http://www.w3.org/TR/xslt, (1999).

[Corby et al. 2004] Corby O., Dieng-Kuntz R. and Faron-Zucker C., Querying the Semantic Web with the CORESE engine. In R. Lopez de Mantaras and L. Saitta eds, Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004), Valencia, Spain, IOS Press, p.705-709, (2004).

[Corby et al. 2006] Corby O., Dieng-Kuntz R., Gandon F. and Faron-Zucker C., Searching the Semantic Web: Approximate Query Processing Based on Ontologies. In IEEE Intelligent Systems Vol.21 No.1 pp. 20-27, (2006).

[Cunningham et al. 2002] Cunningham H., Maynard D., Bontcheva K. and Tablan V., GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. ACL'02, (2002).

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

[Gaizauskas et al. 2003] Gaizauskas R., Demetriou G., Artymiuk P., and Willett P., Bioinformatics applications of information extraction from journal articles. Journal of Bioinformatics, 19(1), pp. 135-143, (2003).

[Gandon et al. 2005] Gandon F., Lo M., Corby O. and Dieng-Kuntz R., Managing enterprise applications as dynamic resources in corporate semantic webs: an application scenario for semantic web services. In W3C Workshop on Frameworks for Semantics in Web Services, http://www.w3.org/2005/04/FSWS/, (2005).

[Golebiowska et al. 2001] Golebiowska J., Dieng-Kuntz R., Corby O. and Mousseau D., Building and Exploiting Ontologies for an Automobile Project Memory. Proc. of the First International Conference on Knowledge Capture (K-CAP), Victoria, October 23–24, (2001).

[Ghoula et al. 2007] Ghoula, N., Khelif, K., and Dieng-Kuntz, R., Supporting patent mining by using ontology-based semantic annotations. In IEEE/WIC/ACM International Conference on Web Intelligence, Silicon Valley, USA (2007).

[Handschuh et al. 2002] Handschuh S., Staab S. and Ciravegna F., S-CREAM – Semi-automatic CREAtion of Metadata. In Gómez-Pérez, A., and Benjamins R. (eds.), Knowledge Engineering and Management: Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain Springer Verlag, LNAI 2473, 358-372 (2002).

[Hoang and Tjoa 2006] Hoang H.H. and Tjoa A., The State of the Art of Ontology-based Query Systems: A Comparison of Existing Approaches, In Proc. of ICOCI06, (2006).

[Humphreys and Lindberg 1993] Humphreys B. and Lindberg D., The UMLS project: making the conceptual connection between users and the information they need. Bulletin of the Medical Library Association 81(2): 170, (1993).

[Kashyap and Borgida 2003] Kashyap V., Borgida A., Representing the UMLS Semantic Network using OWL: (Or “What's in a Semantic Web link?”). In ISWC’2003. Heidelberg: Springer-Verlag; 1-16, (2003).

[Khelif et al. 2005] Khelif K., Dieng-Kuntz R., Barbry P., Semantic web technologies for interpreting DNA microarray analyses: the MEAT system. Proc. of WISE’05, 20-22/11 New York, (2005).

[Khelif 2006] Khelif K., Web sémantique et mémoire d’expériences pour l’analyse du transcriptome. Phd thesis, Nice Sophia Antipolis University, (2006).

[Lassila and Swick 2001] Lassila O. and Swick R., W3C Resource Description framework (RDF) Model and Syntax Specification, http://www.w3.org/TR/REC-rdf-syntax/, (2001).

[Lomax and McCray 2004] Lomax J. and McCray A., Mapping the Gene Ontology into the Unified Medical Language System. Comparative and Functional Genomics, 5:354–361, (2004).

[McBride 2004] McBride B., "RDF Vocabulary Description Language 1.0: RDF Schema”, W3C Recommendation, http://www.w3.org/TR/rdf-schema/, (2004).

[McCray 2003] McCray A., An upper level ontology for the biomedical domain. Comp Functional Genomics; 4: 80-84, (2003).

[McGuinness and Van Harmelen 2004] McGuinness D.L. and Van Harmelen F., OWL Web Ontology Language Overview, http://www.w3.org/TR/owl-features/, (2004).

[Mrabet et al. 2007] Mrabet, Y., Khelif, K., and Dieng-Kuntz, R., Recognising professional-activity groups and web usage mining for web browsing personalisation. In IEEE/WIC/ACM International Conference on Web Intelligence, Silicon Valley, USA (2007).

[Muller et al. 2004] Muller H.M. and Kenny E.E. and Sternberg P.W., Textpresso: an ontology-based information retrieval and extraction system for biological literature PLoS Biol., 2, E309, (2004).

[Nédellec 2002] Nédellec C., Bibliographical Information Extraction in Genomics. IEEE Intelligent Systems & their Applications, N. Shadbolt (ed.), Trends & Controversies - Mining Information for Functional Genomics, 76-80, May-June (2002).

[Ohlbach and Schaffert 2004] Ohlbach H.J. and Schaffert S., eds Workshop on Principles and Practice of Semantic Web Reasoning at the 20th ICLP, St Malo, France, (2004).

[Popov et al. 2004] Popov B., Kiryakov A., Ognyanoff D., Manov D. and Kirilov A., KIM – a semantic annotation platform for information extraction and retrieval. Natural Language Engineering, 10, Issues 3-4, 375-392, (2004).

[Rector et al. 1996] Rector A., Rogers J.E. and Pole P., The GALEN High Level Ontology. Fourteenth International Congress of the European Federation for Medical Informatics, MIE96, Copenhagen, Denmark, (1996).

[Rzhetsky et al. 2004] Rzhetsky A., Iossifov I., Koike T., Krauthammer M., Kra P., GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inf 37: 43–53, (2004).

[Séguéla and Aussenac-Gilles 1999] Séguéla P. and Aussenac-Gilles N., Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In Proc. of IC'99, Paris, 79-88, (1999).

[Schmid 1994] Schmid H., Probabilistic part-of-speech tagging using decision trees. In proceedings of International Conference on New Methods in Language Processing. Manchester, (1994).

[Shatkay et al. 2002] Shatkay H., Edwards S., and Boguski M., Information Retrieval Meets Gene Analysis. IEEE Intelligent Systems 17, 2, 45-53, (2002).

[Shatkay and Feldman 2003] Shatkay H. and Feldman R., Mining the biomedical literature in the genomic era: an overview, Journal of Computational Biology, 10, 821–855, (2003).

[Staab 2002] Staab S., Mining Information for Functional Genomics. IEEE Intelligent Systems & their Applications, March-April, 66-80, (2002).

[Uren et al., 2006] Uren V., Cimiano P., Iria J., Handschuh S., Vargas-Vera M., Motta E. et Ciravegna F., Semantic annotation for knowledge management: Requirements and a survey of the state of the art. In Web Semantics, Volume 4, Issue 1, 14-28, 2006.

[Vargas-Vera et al. 2002] Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A. and Ciravegna F., MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup, In Gómez-Pérez A. and Benjamins R. eds, Knowledge Engineering and Management: Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain, Springer Verlag LNAI 2473, 379-391, (2002).

[W3C 2004] Summary Report - W3C Workshop on Semantic Web for Life Sciences. http://www.w3.org/2004/10/swls-workshop-report.html, (2004).

[Zweigenbaum 1994] Zweigenbaum P., MENELAS: an access system for medical records using natural language. Comput Methods Programs Biomed. Oct;45(1-2):117-20, (1994).