Semantic Web Technologies for Analysis of Transcriptome

Rose Dieng-Kuntz[1], Khaled Khelif[1], Olivier Corby[1], Pascal Barbry[2]

[1]INRIA, UR Sophia Antipolis project ACACIA
2004, route des lucioles BP93, 06902 Sophia Antipolis Cedex, France
Tel : (33) - (0)4 92 38 78 10, Fax : (33) – (0)4 92 38 77 83
email: {Rose.Dieng, Khaled.Khelif, Olivier.Corby}@sophia.inria.fr

[2]IPMC, Institut de Pharmacologie Moléculaire et Cellulaire
660 Route des Lucioles, 06560 Valbonne Sophia-Antipolis
email : barbry@ipmc.cnrs.fr, tél : 04 93 95 77 77 fax : 04 93 95 77 08

1. Organizational Semantic Webs

The Acacia team studies knowledge management through the building of an organizational memory, that we propose to materialize an organizational memory through an "organizational semantic web" constituted of:

- resources : they can be documents (in various formats such as XML, HTML, or even classic formats), but these resources can also correspond to people, services, software or programs,
- ontologies (describing the conceptual vocabulary shared by one or several communities in the oraganisation),
- semantic annotations on these resources (i.e. contents of documents, skills of persons or characteristics of services / software / programs), based on these ontologies,
- with diffusion on the Intranet or the corporate Web.

The Acacia team studies the research topics linked to the creation and evolution of the components of a corporate semantic Web (resources, ontologies, annotations) from heterogeneous knowledge sources (in particular, textual corpora and databases), with possibly multiple viewpoints, and the us of ontologies and semantic annotations by a semantic search engine having inference capabilities for offering ontology-guided information retrieval.
IPMC offers a national platform for creation and analysis of DNA microarrays, and is responsible of the MEDIANTE project for production of pangenomic biochips for human and for mouse.

This paper presents a new application of semantic web in the context of life sciences: the MEAT (Mémoire d'Expériences pour l'Analyse du Transcriptome) – Experiment Memory for Transcriptome Analysis. This collaborative project aims at supporting a community of biologists by building a memory of experiments in DNA microarrays: it constitutes a specific case of organizational semantic web, at the scale of a community.
After describing the whole project, we will conclude on our needs from Semantic Web technology viewpoint.

2. The MEAT project for Transcriptome analysis

2.1 Requirements from biologists' viewpoint

The needs of the biologist working on biochips, can be summarized as follows.

1. *Biochip Production control and Archiving of Experimental Data*.
2. *Analysis of Expression Data.*
3. *Support to validation of experimental results.* Once the biological experiment carried out and after the analysis of the data, only a part of the modifications observed (i.e. induced or repressed genes) can be explained by the model postulated initially by the biologist. The biologist's concern is thus to connect the new experimental data generated by the experiment

to the information contained in reference sites such as NCBI sites (LocusLink, UniGene, OMIM, HomoloGene, GEO, PubMed...), or more specific sites, producing information on the specific expression of genes in such or such tissue (http://www.ncbi.nlm.nih.gov/SAGE; Osprey...). During this activity, the biologist searches for information on a given site (RefSeq, GO, etc), by using the adequate name of the gene s/he is interested in, and s/he retrieves in a more or less systematic way all information relating to studied gene. The goal is to establish what was published on a given gene, on its possible interactions with other genes, its contribution to a molecular function, etc... Once found, this kind of information is used to validate the results of experiments, i.e. to define the expression profiles coherent with the private data (information retrieval from the base of experiments) or with the literature (information retrieval from distant documentation data bases).

4. *Support to interpretation of the experiment results.* In the experimental results, the biologist tries to identify not yet known relations between genes (same behaviour, participation in the same phenomenon or a phenomenon related to the principal phenomenon, same biological functions, role in a pathology etc.) and relations between experiments (validation or not of later experiments, etc). Once the experiments carried out, some genes are identified as activated or inhibited. This process enables to answer the following questions:
   - which new probes would be interesting to be added in order to complete the study (proteins in "interaction" with genes affected by the treatment, defective probes...)?
   - which other experiments not involving the biochip technology would be necessary to reinforce the conclusions of the experiments carried out?
   - what are the conclusions of the experiment in progress?

2.2 Towards a DNA-microarray experiment memory

Taking into account such needs, we aim at building an experiment memory for transcriptome analysis with data stored via the Web by the technicians of the biochip platform (for slide production) and by the users of the biochips (for the hybridizations). The approach proposed consists of the following stages:

1. Checking and validation by the user biologist of the probes available on the biochip, and assisted selection of a relevant subset for the project.
2. Order slides to launch a new biochip experiment.
3. Meanwhile, submission of journal articles, concerning genes supposed a priori interesting for the future experiment or for the biological phenomena studied.
4. Constitution of an electronic document corpus from these articles.
5. Creation of semantic annotations on these articles.
6. After realization of the experiment, storage of its description and of its results in MEDIANTE according to the interchange format defined by Array Express (http://www.ebi.ac.uk/arrayexpress/). The users will be able to deposit their experiments on the EBI server.
7. Statistical analysis of results: the biologist will use data mining tools for mining either the data created specially for MEDIANTE (MEAT-Miner) or the data resulting from already existing projects, such as BioConductor (http://www.bioconductor.org).
8. Interpretation of the results, which may imply additional bibliographical searches guided by ontologies, in order to confirm or to infirm the biologist's hypotheses.
9. Addition by the user of new annotations on the experiment, thanks to the annotation editor.

These various stages will be carried out using the four following components (cf. figure 1):

1. the MEDIANTE storage interface, using PostGreSQL databases.
2. the module of annotation and search on the experiment memory (MEAT-Annot&Search), which relies on semantic Web technologies,
3. the data mining module (MEAT-Miner), which implements knowledge discovery techniques (segmentation, classification, prediction...),

4. the module of ontological mediator (MEAT-Onto).

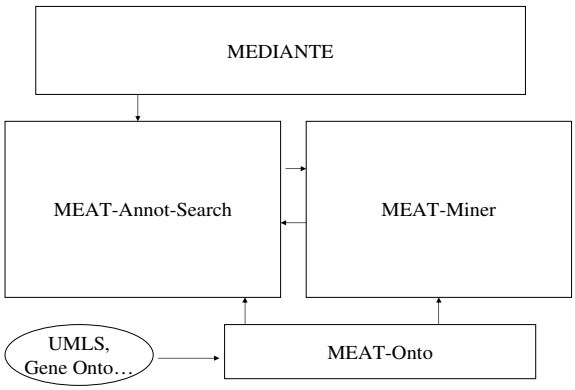

Figure 1: Global architecture of MEDIANTE-MAT

Most of the storage interface is already in production (http://medcal.ipmc.cnrs.fr:8080/mediante) at IPMC. The modules MEAT-Miner and MEAT-Onto will be developed by our partners from I3S. Therefore, in this paper, we will detail only the MEAT-Annot&Search module, developed by the ACACIA team. This module will include the following components:

- The *tool for annotation acquisition* is composed of a manual annotation editor and of a tool for semi-automatic generation of annotations from a textual corpus. These annotations consist of (a) specific instances of concepts of the ontology (genes, biological processes, cellular components, body parts, pathologies, etc.) or (b) semantic relations between these concepts (e.g. interactions between genes, relations between genes and pathologies, direct or indirect participation of some genes in some biological phenomena, behaviors of these genes with respect to chemical treatments etc).
  - The *manual annotation editor* will be guided by one or several biomedical ontologies (Gene Ontology, UMLS) provided via MEAT-Onto. It will enable to associate to an element of a particular document (text or image) an annotation translated into RDF.
  - The *generator of annotations from textual corpus* has already been implemented (Khelif et al, 2004). Taking as input the textual corpus, it relies on various Natural Language Processing (NLP) tools: existing tools such as the term and relation extractor, Syntex, and the tool GATE (General for Text Engineering Structures) (Cunningham et al, 2002, 2003), with our own extensions dedicated to extraction of semantic relations involving genes. We consider the UMLS semantic network associated to the UMLS meta-thesaurus, as an ontology transformed by a script into RDF(S). The annotation generator is guided both by this ontology and by a relation extraction grammar manually created after extraction by Syntex of the verbal syntagms appearing in a sample textual corpus of scientific articles in biology and representative of the relations interesting in the biology field. MEAT-Annot enables to extract from texts the terms corresponding to UMLS concepts, and to locate in the text some semantic relations between them (e.g. interactions between genes or interactions between genes and other entities). The annotation generator also has a graphical interface enabling the expert to validate the annotations carried out.
  - The *base of annotations* structured in several bases dedicated to the various biochip projects, with on the one hand, annotations concerning the reference articles of these projects and, on the other hand, the annotations concerning the results of these projects. General knowledge in biology, non associated to a particular document but likely to be used for inferences on the annotations during the phase of search for information. These annotations should even be contextual and depend on the point of view adopted by the

annotator: so the system must be able to manage some contradictory annotations relying on different points of view.

- The *Corese semantic search engine* (Corby et al, 2000, 2002, 2004) developed by the Acacia team will enable, via a MEAT-dedicated query interface, to rely on the ontology and on the annotation base possibly completed by the inferences using a rule base, for offering relevant answers to the biologists' requests. We will develop extensions of Corese and in particular graphic interfaces dedicated to MEAT: a query interface (based on the queries privileged by the biologists searching scientific articles enabling to help them to interpret the experiment results or to validate their interpretation), a base of inference rules dedicated to MEAT and an interface for presentation of answers and for navigation in such anwers.

Remark: In addition to ontology population and annotation generation, the linguistic techniques used for the semi-automatic generation of the semantic annotations (Khelif et al, 2004) could also be used for enrichment of an existing ontology (e.g. UMLS, Gene Ontology) or even for creation of an ontology dedicated to the experiments biochips.

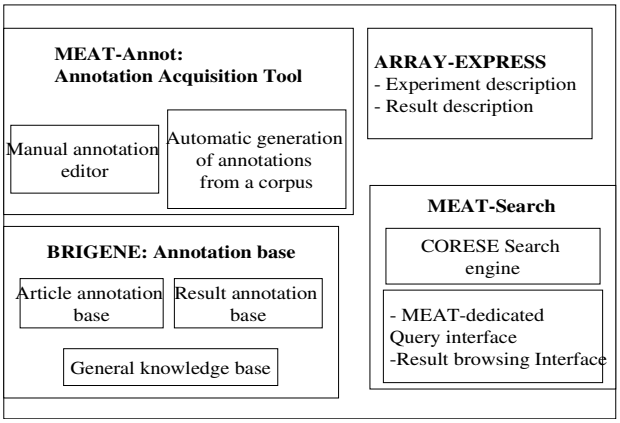

Figure 2: Architecture of MEAT-Annot&Search

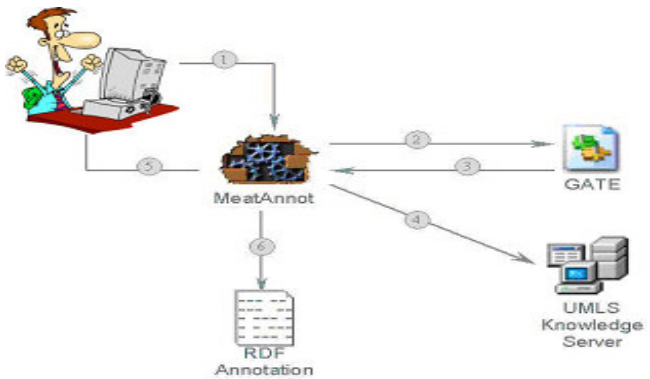

Figure 3. Interactions with MEAT-Annot

Figure 3 describes the interactions with the MeatAnnot system:
(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 sends them to UMLS Knowledge Server for check their existence in UMLS and for obtaining their semantic types and their synonym terms,
(5) MeatAnnot presents extracted information to the user for validation, through a graphical interface,
(6) After validation by the user, MeatAnnot generates the RDF annotation associated to the studied article and stores it in the article annotation base.

Our test corpus was a set of articles related to lung diseases provided by the IPMC team working on DNA-microarray experiments. This test phase enabled: (i) the validation of the concept instantiation method, (ii) the validation of the relation extraction method and (iii) the verification of the coherence and consistency of the generated annotations, this verification being performed using the RDF-dedicated semantic search engine CORESE developed in the Acacia team.

Conclusions and discussion

The paper presented our approach on organizational semantic webs, illustrated through the MEAT project aimed at facilitating validation and interpretation of DNA-microarray experiments results. Our corporate semantic web approach is used in another application in life sciences: the Life Line project (Dieng-Kuntz et al, 2004), aimed at supporting cooperative work in a healthcare network, relying on building semi-automatically a medical ontology from heterogeneous sources (database, textual corpus) and on the use of this ontology for a cooperative tool.
What are the requirements of MEAT project from semantic web technology viewpoint?

- We rely on RDF for expressing annotations, on RDFS for representing ontologies and we handle them through the RDF-dedicated search engine, Corese. But as more and more ontologies in biomedical domain will evolve towards OWL, we will need to adapt Corese to OWL.
- For queries about the annotation base, we rely on Corese query language (Corby et al, 2004): it would be interesting to compare it to the W3C-recommended query language.
- For rules enabling to complete the annotation base through inferences, we rely on Corese rule language (Corby et al, 2002, 2004), that could also be compared with related work on rule languages aimed at Semantic Web (RuleML, etc.).
- For taking into account contextual annotations, we would need extensions of RDF so as to enable to express multiple contexts or multiple viewpoints. Moreover, such annotations could even be related to several ontologies, with an explicit alignment between such ontologies.
- For querying the base of past biochip experiments, it could be useful to express temporal queries. For tackling the possible evolution of ontologies or annotations through time, we also need to express, store and reason on temporally evolving ontologies or annotations.
- If instead of relying on reference articles provided by the launchers of a DNA experiment, MEAT-Annot had to perform scientific watch on the whole open Web, we would need to solve a huge scalability problem.
- One open question is whether the semi-automation of annotation creation may require deeper NLP techniques than those presently offered by NLP tools.

Acknowledgements

We thank very much the IPMC team working on DNA-microarrays, Remy Bars from Bayers Cropscience, Didier Bourigault for providing us the results of Syntex on our corpus, Laurent Alamarguy for his assistance in linguistic domain, Martine Collard, Nhanh le Thanh and Ricardo Martinez that work on MEAT-Miner and MEAT-Onto and PACA region which funds this work by a regional grant.

References

Ashburner M. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25 .

Berners-Lee T., Hendler J. & Lassila O., The Semantic Web, Scientific American, 84(5) p. 34-43. (2001)

Blaschke C. & Valencia A., (2002). Molecular biology nomenclature thwarts information-extraction progress. IEEE Intelligent Systems & their Applications, p. 73-76.

Blaschke C, Andrade MA, Ouzounis C and Valencia A. (1999). Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. ISMB99, 60-67.

Blaschke C, Oliveros JC and Valencia A. (2001). Mining functional information associated to expression arrays. Functional and Integrative Genomics 4: 256-268.

Blaschke C and Valencia A. (2001). Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. Comp. Funct. Genom. 2: 196-206.

Blaschke C, Hoffmann R, Oliveros JC and Valencia A. (2001). Extracting information automatically from the biolobical literature. Comp. Funct. Genom. 2: 310-313.

Blaschke C and Valencia A. (2001). The potential use of SUISEKI as a protein interaction discovery tool, Genome Informatics Series 12: 123-134.

Corby, O., Dieng, R., Hébert, C. (2000). A Conceptual Graph Model for W3C Resource Description Framework. In B. Ganter, G. W. Mineau eds, Conceptual Structures: Theory, Tools, and Applications, Proc. of ICCS'2000, Springer-Verlag, LNAI 1867, Darmstadt, Germany, August 13-17, p. 468-482.

Corby, O., Faron-Zucker, C.: Corese (2002). A Corporate Semantic Web Engine. In Workshop on Real World RDF and Semantic Web Applications - 11th International World Wide Web Conference 2002 Hawaii. May 2002. http://paul.rutgers.edu/~kashyap/workshop.html

Corby, O., Dieng-Kuntz, R., Faron-Zucker, C. (2004). Querying the Semantic Web with the CORESE Search Engine. Proc. of the 16th European Conference on Artificial Intelligence (ECAI'2004), Valencia (Spain), August 25-27th, 2004.

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

Cunningham H., Maynard D., Bontcheva K., Tablan V., Ursu C. & Dimitrov M. (2003). Developing Language Processing Components with GATE (User Guide).

Dieng-Kuntz, R., Minier, D., Corby, F., Ruzicka, M., Corby, O., Alamarguy, L., Luong, P (2004). Medical Ontology and Virtual Staff for a Health Network, In E. Motta et al, eds, Engineering Knowledge in the Age of the Semantic Web, 14th Int. Conference, EKAW'2004, Whittlebury Hall, UK, October 5-8th, 2004, p. 187-202.

Kashyap V., Borgida A. (2003). Representing the UMLS Semantic Network Using OWL: (Or "What's in a Semantic Web Link?"). ISWC 2003, USA. p 1-16.

Khelif K., Dieng-Kuntz R. (2004) - - Annotations sémantiques pour le domaine des biopuces, Actes des 15èmes journées francophones d'Ingénierie des Connaissances (IC'2004), Lyon, 4-6 mai 2004, PUG, p. 273-284.

Khelif K., Dieng-Kuntz R. (2004) - Ontology-Based Semantic Annotations for Biochip Domain, Proc. ECAI'2004 workshop on Knowledge Management and Organizational Memories, Valencia, August 2004 and Proc. of EKAW'2004 Workshop on EKAW 2004 Workshop on the Application of Language and Semantic Technologies to support Knowledge Management Processes, U.K., October 2004.

Kim S., Alani H., Hall W., Lewis P., Millard D., Shadbolt N. and Weal M. (2002). Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web. In Proceedings of Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM'02), pages pp. 1-6, Lyon.

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

Lindberg D., Humphreys B., Mccray A. (1993). The Unified Medical Language System. Methods Inf Med, p 281-291,

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

National Library of Medicine (2003). UMLS Knowledge Source. 14th Edition, Jan. 2003 Doc. National Institute of Health – National Library of Medicine, Bethesda, Md, USA.

Proux, D., et al. (2000). A Pragmatic Information Extraction Strategy for gathering Data on Genetic Interaction, in proceedings of ISMB.

Schulze-Kremer S., Smith B., Kumar A. (2002). Revising the UMLS Semantic Network

Shatkay H., Edwards S. & Boguski M. (2002). Information Retrieval Meets Gene Analysis. IEEE Intelligent Systems & their Applications, p. 45-53.

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