A semantic collaborative system for the management of translational research projects

Matteo Gabetta[1]✉, Giuseppe Milani[1], Cristiana Larizza[1], Valentina Favalli[2], Eloisa Arbustini[2], Riccardo Bellazzi[1]

[1]Dipartimento di Ingegneria Industriale e dell'Informazione, University of Pavia, Pavia, Italy
[2]IRCCS Fondazione Policlinico S. Matteo, Pavia Italy

Motivation and Objectives

Translational research projects aim at combining -omics, structural and functional studies with clinical investigation results to translate basic knowledge of genetic diseases into routine clinical practice. Biomedical informatics can fruitfully support this kind of research by implementing information technology solutions to support the multidisciplinary project team in the different phases of its investigation.

In this paper we present a semantic wiki-based system purposely implemented for supporting the consortium members of the EU project Inheritance in sharing and disseminating data and knowledge about genetic dilated cardiomyopathies (DCM) [Ahamad et al, 2005]. It consists of a collaborative system that is used to track project activities, share ideas and data, foster exchange of information between the investigators to support several activities of the INHERITANCE translational research project. Moreover, it can be used to easily manage the scientific research products by adding semantic tags on the basis of the underlying knowledge model. A Natural Language Processing (NLP) based module has been developed to this aim; it extracts the relevant molecular and medical concepts from the scientific material shared by the project team and store them as RDF form by enabling the semantic querying of data

Methods

The INHERITANCE Project's Semantic Wiki has been designed and implemented for two purposes: to manage in a collaborative and fast shareable way information and documents related to the organizational aspects of the project and to allow users to share scientific documents automatically analysed and annotated thanks to an integrated NLP based tool.

To build such a Wiki we choose to extend the standard MediaWiki [web site: http://www.mediawiki.org/wiki/MediaWiki], last accessed on July 27, 2012) platform with its most popular semantic extension, called Semantic MediaWiki [Krotzsch et al, 2006].

The first step of the environment setup consisted of defining the Categories necessary to model the information managed inside the Wiki, and the Templates and Forms, which are required to define the content of each category.

In the first release of the Wiki we have implemented the "Person", "Organization", "Meeting" and "Work Package" Categories to represent the organizational aspects of the project, and the "Protein", "Gene" and "Dilated Cardiomyopathy Documents" Categories to model the scientific aspects.

In the typical system use case the authorized users manually insert the organizational data using the proper Templates and Forms; these information will be available for any further interrogation with the smart querying tools available in the Wiki. The main reason for not implementing an automatic import process of these data from the project material is their actual nature: indeed they are spread among many different documents, but their relatively small number doesn't justify the presence of an automatic extraction tool.

Differently, the scientific knowledge management section of the Wiki is designed to deal with an arbitrary large number of documents; therefore we implemented, on top of the Wiki, a concept extraction system able to: a) let the user upload a document (in plain text, pdf or MS Word format) and choose the name of the Wiki page where the document will be stored; b) extract genes and proteins cited inside the document, recursively checking if the gene/protein is already present in the Wiki (otherwise a page for the new gene/protein is created) and link these pages to the one containing the document; c) add the page representing the document to the Wiki.

To realize such a solution we designed a servlet directly accessible from a special page of the Wiki called "NLP"; the concept extraction module

of the servlet is based on Gate [H. Cunningham, 2002], an open-source library for natural language processing. This tool combines a standard (and already implemented) text analysis pipeline with some modules purposely developed in order to extract the cited genes (exploiting the Entrez Gene NCBI's database [Maglott et al, 2005]) and proteins (exploiting Uniprot [The UniProt Consortium, 2012]).

In addition, when a new page representing a gene or a protein is created, the system, thanks to the NCBI Entrez Programming Utilities tools [web site: http://www.ncbi.nlm.nih.gov/books/NBK25500/] (last accessed on July 27, 2012), automatically associates to the page the five most recent articles from Pubmed that have that gene/protein as topic.

Once the Wiki has been populated with the project's data, it is possible to perform, beyond all the standard tasks of a traditional Wiki (update, content modification, old pages restore, discussion, etc.), also some smart querying operations that exploit the semantic nature of the data. The semantic query tools available in the Wiki use two distinct languages: a simple query language, to perform queries within the Wiki's data, and SPARQL [Herman, 2008] that is the standard query language for the semantic web, opening the Wiki to the possibility of a future integration with many other available repositories of linked data [web site: http://linkeddata.org/] (last accessed on July 27, 2012).

Results and Discussion

Actually, the INHERITANCE semantic wiki is up and running at the URL http://www.labmedinfo.org:8123/mediawiki/index.php/Main_Page and is made available to all the consortium members to track the project activities (meetings, partners, work packages) and manage every product of the project (deliverables, scientific papers). A Summary page has been defined to synthetize all the project activities and participants information. Moreover, the RelFinder browser [http://www.visualdataweb.org/relfinder.php] (last accessed on July 30, 2012), useful to look for rela-

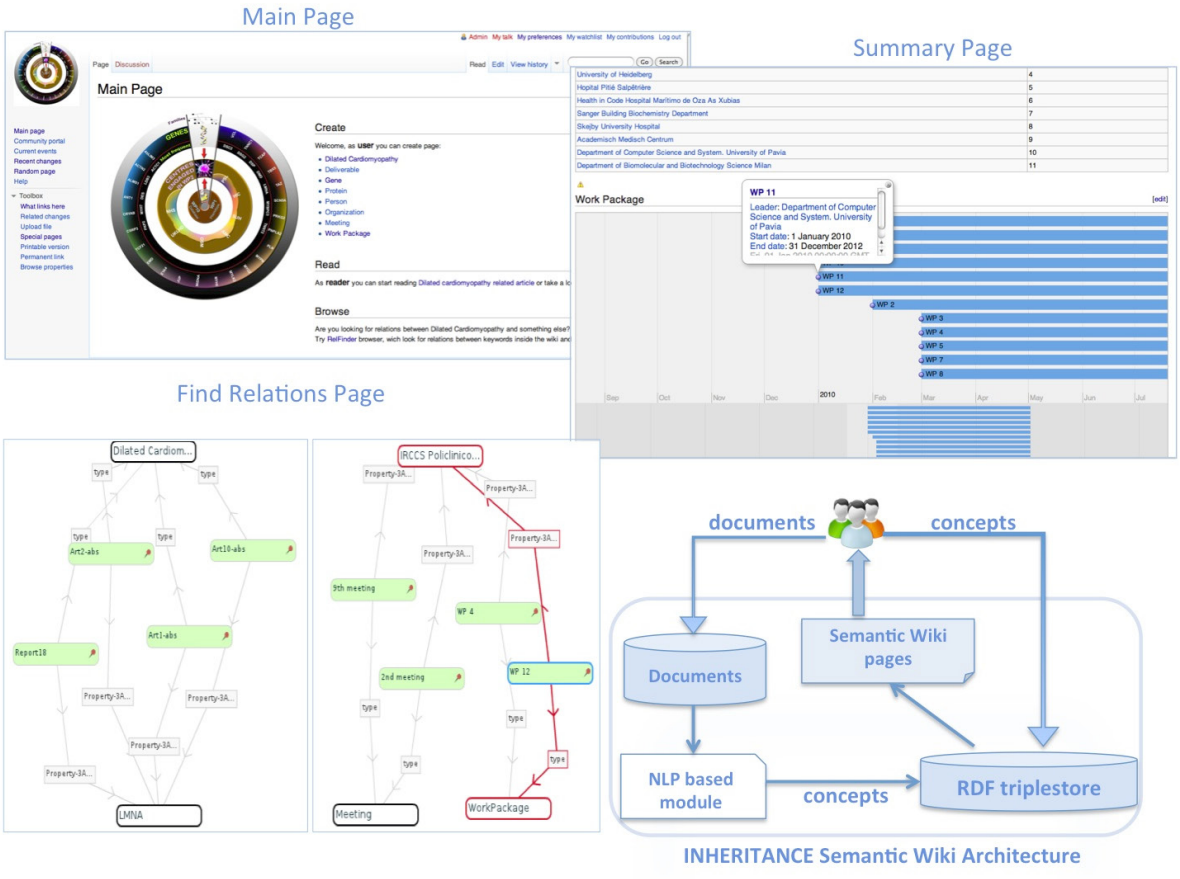

Figure 1 – The Semantic Wiki architecture and the Main, Summary and Find Relations project pages screenshots

tions between keywords inside the wiki and show the relations graph (eg. Person- Organization or Meeting-Organization relations), has been made available (Figure 1).

Currently the main goal of the semantic wiki is to support the INHERITANCE research group from two distinct points of view: the organizational and the scientific data management and sharing. While all the features related to the organizational aspects have been developed and tested by the users, the scientific knowledge management section of the wiki is still under development. The current prototype provides some basic features such as the scientific documents storage and mapping to custom categories, the NLP facilities for data extraction and the automatic linkage to relevant scientific literature. Nonetheless the upgrade of the system with new tools (e.g. link to specific DCM resources and integration with biological databases) doesn't entail relevant technical problem, and its actual implementation, although planned, depends on the future developments of the INHERITANCE project and on the users' feedback after the system evaluation.

At this moment the NLP based module has been used to annotate 10 documents and extract 13 genes and 10 proteins. In future we plan to link the data to external resources from across the Linked Data community.

Acknowledgements

This work is part of the INHERITANCE Project, funded by the European Commission.

References

1. Ahamad F, Seidman JG, Seidman CE. (2005) The genetic basis of cardiac remodelling. Annu Rev Genomics. Hum Genet 6, 185. doi: 10.1146/annurev.genom.6.080604.162132
2. H. Cunningham, D. Maynard, et al (2002) GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia.
3. Maglott D, Ostell J, Pruitt KD, Tatusova T. (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res, 33 (Database Issue):D54-8
4. Herman, W3C Semantic Web Activity News - SPARQL is a Recommendation, http://www.w3.org/blog/SW/2008/01/15/sparql _ is _ a _ recommendation/ W3.org. 2008-01-15. (Last accessed on July 27, 2012)
5. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res., 40, D71–D75.
6. Krotzsch, M., Vrandecic, D. and Volkel, M. 2006. Semantic MediaWiki. Proceedings of the Fifth International Se- mantic Web Conference, pp 935-942, Springer, November 2006.