DC-THERA Directory, a Knowledge Management System for the support of the European Dendritic Cell Immunology Community

Marco Brandizi[1,*], Michaela Gündel[1,*], Ciro Scognamiglio[1], Andrea Splendiani[1]

[1] Leaf Bioscience s.r.l., Via G. Puccini 3, 20121 Milan, Italy
{marco, michaela.guendel, ciro.scognamiglio, andrea}@leafbioscience.com
[*] These authors equally contributed to this work.

Abstract. DC-THERA Directory is a web portal to support collaboration, communication and knowledge sharing within DC-THERA, a community focused on immunology. We show how we have faced the problem of representing and managing highly heterogeneous and interconnected knowledge. One aspect of the application interface is the search and navigation through web ontologies. Another aspect is the dynamic representation of information entities having variable sets of properties. These results have been achieved by adopting a modelling approach that combines traditional object-oriented modelling with a triple-based knowledge representation. We discuss advantages of such an approach, especially for what concerns the possible future integration of other information sources.

Keywords: Knowledge Management Systems, Bio-Ontologies, Dendritic Cells, On Line Communities, Semantic Web.

Introduction

Knowledge Management Systems based on web technologies are tools widely used for promoting collaboration and information flow among organizations, including scientific communities [1]. Moreover, standard terminology and formal ontologies in knowledge applications are particularly important for the life sciences domain [2], as this is characterised by highly heterogeneous and interconnected information. These technologies are also useful for supporting cooperation in communities put together by research projects [3,4]. An example of that is DC-THERA[1], a European project, focusing on integrating a large community of researchers and therapists on dendritic cell research, an important topic in immunology.

In the DC-THERA context, people often need to answer questions like: are certain entities (e.g., blood samples of a certain kind or a purification method) available in the network? Who is maintaining them? Who is expert in their usage? How can I contact them? These requirements are addressed in the DC-THERA Directory, a web-based knowledge management system that allows to collect summary information about the bio-medical resources available among the DC-THERA researchers. The Directory eases knowledge sharing and exchange by providing a single, unified access point to a variety of bio-entities. It promotes collaboration by representing links between bio-resources and people who are related to them (for example, someone having experience in applying a given laboratory protocol, or someone else studying a certain data set).

1 http://www.dc-thera.org

Furthermore, by integrating other biological repositories (e.g., about specific experimental data, or about scientific publications), it allows to narrow down an initial search or find desired details that are managed outside the scope of the Directory itself. Finally, having an electronic repository where to store reference information about the work and achievements produced in the context of an EU-funded research project helps in not dispersing such information and potentially keeping it available to the general public.

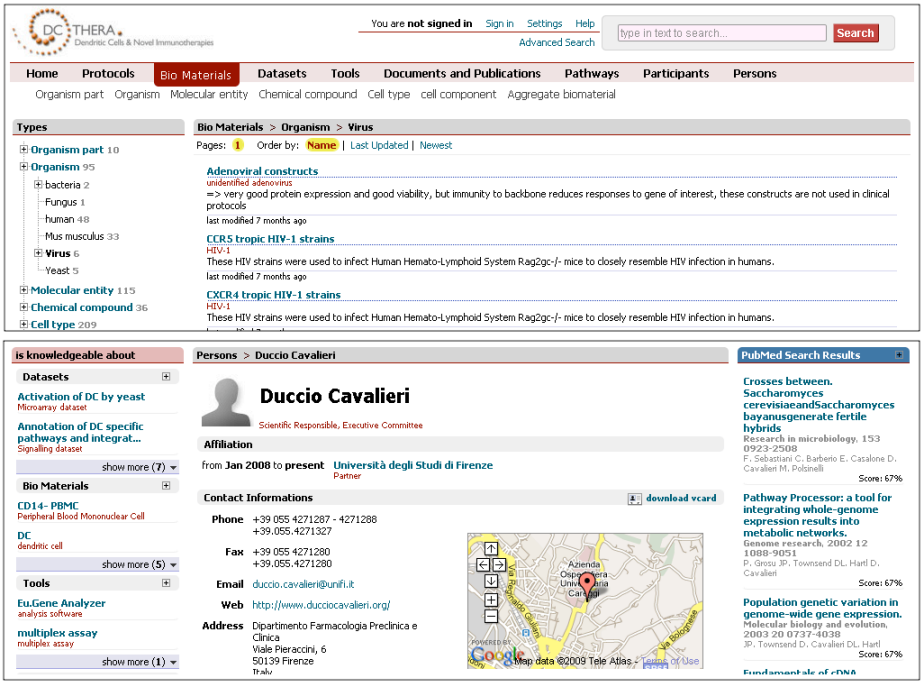

Figure 1: a list of resources (top) and a person description.

The Directory

The application currently supports the main categories visible in fig. 1. Clicking on one of them, or their sub-categories, a list of existing resources is shown, together with a taxonomy of sub-types defined for that category, which is built based on the underlying ontologies. From here, it is possible to open the details form of a particular resource in a "subject centric" view, that puts together resource properties, such as title or description, and links to other items (fig. 1, bottom). Both the details view and the list of properties can flexibly vary, depending on the specific resource.

Another use case is the classical keyword-based search. An auto-completion feature suggests search hints while typing, by dynamically querying (via AJAX) the knowledge base and proposing expressions such as term variants, associated keywords and synonyms (again, the knowledge contained in the ontologies is used for that). Results are presented by highlighting the terms in the text. Depending on the re-

source category that is queried, additional related external data are shown (e.g.: list of publications in fig. 1; gene expression information, coming from EBI's Gene Expression Atlas[2] or BASE installations [5]).

Architecture and Implementation. DC-THERA Directory has been designed by adopting a modelling approach that is similar to the one proposed in [6]. On the one hand, an object model has been defined, implemented in PHP (fig. 2), containing classes for the main resource types and defining general and specific properties in top-level and pertinent classes respectively. On the other hand, the representation of property/value pairs and helpers for querying the knowledge base by using triple patterns has been embedded in the object model (e.g.: *<r, p, *>* finds all resources that are p-related to r).

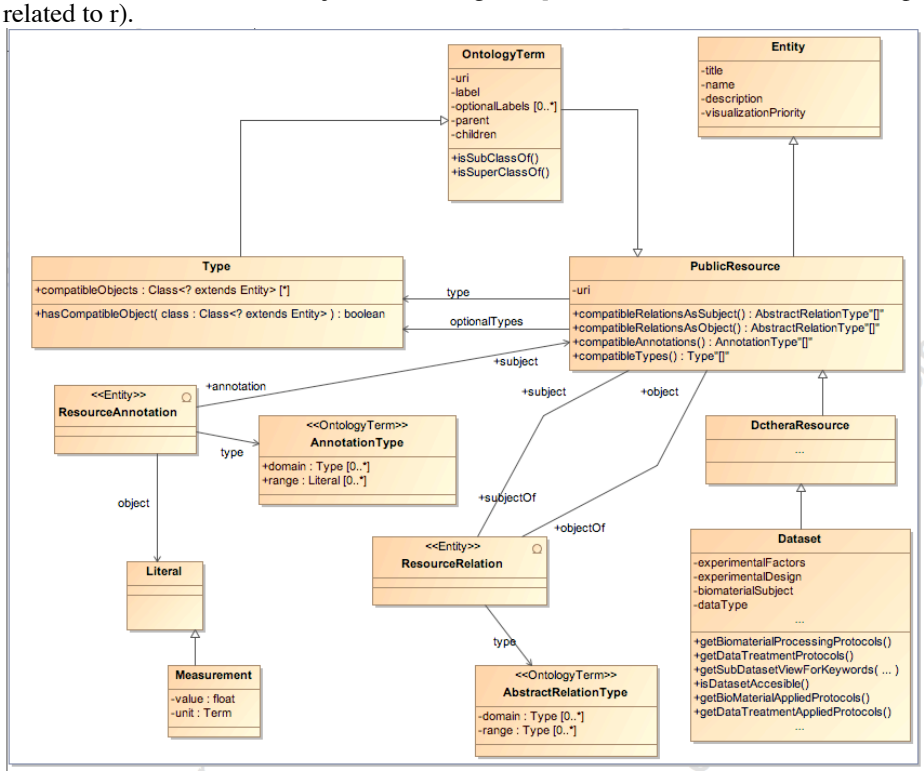

Figure 2: the DC-THERA Directory object model.

This "mixed" model makes use of the advantages of both the world of structured object-oriented architecture and the world of semi-structured "RDF-like" triples. The PHP classes can contain implementation-specific code, such as methods invoked by the Symfony[3] framework (e.g.: the Dataset class in fig. 2). The "triplified" part allows to dynamically define any set of properties and links for a resource, e.g., characterising the resource with terms from external ontologies, or embedding triple statements obtained by external services. For example, we have selected and imported those Open Biomedical Ontologies [7] that are suitable to represent the kind of information

2 http://www.ebi.ac.uk/gxa
3 http://www.symfony-project.org

currently available in the Directory and the tackled domain – especially OBI, the Ontology for Biomedical Investigations[4].

Discussion and future work

A first stable version of the Directory was released to the project's participants in July 2009. Before, a test had been done with a selected panel of pilot users, who helped in evaluating and further improving the quality of the tool and its contents by answering a questionnaire (unpublished). This had confirmed that the application and the way it has been developed is found useful by the users. As an example of improvement suggested by the survey, we are using the BioLexicon [8] text mining tool to extend ontology classes with term variants, extracted from literature text mining. This will improve keyword-based searches, which already use ontologies for aspects like synonyms or semantically close terms.

Another feature under development is an export of the Directory content to the RDF/OWL format. This could, for instance, be useful to integrate and compare the knowledge base with similar on line resources, to find useful information without the need to query many repositories one by one [9]. Another example is comparing the Directory by semantic similarity, analysing ontologically related terms [10].

In conclusion, as the first user feedback suggests, the DC-THERA Directory provides an effective way to make available the knowledge asset developed in the five-year project time. As far as we know, it is the first time that the hereby described hybrid modelling approach is used to build a rapid web development infrastructure. This allows both the development of similar bio-medical applications and, thanks to the fact it is compatible with Semantic Web technologies, to integrate them together.

References

1. Das, S., et al.: Building biomedical web communities using a semantically aware content management system. Brief Bioinform;10(2):129-38 (2009)
2. Coskun, G., et al.: Towards Corporate Semantic Web: Requirements and Use Cases. Freie Universität Berlin, Tech Rep TR-B-08-09 (2008)
3. Clark, T., Kinoshita, J.: Alzforum and SWAN: The Present and Future of Scientific Web Communities. Brief. in Bioinformatics 8(3):163-171 (2007)
4. Gaines, B. R., Shaw, M. L. G.: Knowledge management for research communities. Proc AAAI Spring Symposium on A.I. in Knowledge Management, Stanford University, pp 55-62 (1997)
5. Saal, L. H., et al.: BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data. Genome Biology 3(8): software0003.1-0003.6 (2002)
6. Puleston, C., et al.: Integrating object-oriented and ontological representations: A case study in Java and OWL. The Semantic Web - ISWC 2008, pp. 130-145 (2008)
7. Smith, B., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. (11):1251-5 (2007)
8. Rebholz-Schuhmann, D., et al.: BioLexicon: Towards a Reference Terminological Resource in the Biomedical Domain. Proc ISMB-2008 (2008)
9. Heng, T. S. P., et al.: The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunology 9, 1091 - 1094 (2008)
10. Pedersen, T., et al.: Measures of semantic similarity and relatedness in the biomedical domain. J of Biomedical Informatics, Vol. 40, Issue 3. (2007)

4 http://purl.obofoundry.org/obo/obi