Practical Experiences in Building Ontology-based Retrieval Systems

Elena Paslaru Bontas

Freie Universität Berlin
Institut für Informatik
Takustr. 9, D-14195 Berlin, Germany
paslaru@inf.fu-berlin.de

Abstract. This paper describes practical experiences and lessons learned in the project "A Semantic Web for Pathology", a project using domain ontologies and ontology-driven natural language processing to support a content-based retrieval of text and image-based medical information. In trying to develop the target application ontology we investigated the potential of reusing the huge amount of domain knowledge already available in ontology-like form in the medical domain as an input for the domain conceptualization. According to our experiences we confirm previous findings in the knowledge acquisition literature and in recent surveys of the state of the art in the ontology engineering area: 1). building ontology-based applications is still a tedious process mainly because of the lack of mature tools and methods, which can handle the requirements of real-world applications; and 2). using existing ontologies in new application contexts is currently related to considerable efforts, which might outweigh the benefits of the reuse process. We use the insights gained during this project to derive a set of guidelines for developing Semantic Web retrieval applications in similar domains.

1 Introduction

Parallel with the continuous dissemination of ontologies and the emergence of methods and tools which assist the ontology developers in creating, managing and extending them, ontology engineering is currently evolving from a pure research topic to real-world applications. This state of the art is emphasized by the wide range of European and international projects with major industry involvement and by the increasing interest of small and medium size enterprizes asking for consultancy in this domain. However, due to the well-known difficulties associated with ontology engineering activities, building and deploying ontologies at a large scale, beyond the boundaries of the academic community, has to be supported not only by elaborated technologies and tools, but also by *extended case studies and comprehensive guidelines*, which aid the engineering team in practical situations in their attempt to develop ontology-based applications. This need is currently addressed in several joined initiatives, such as the W3C Semantic Web Best Practices and Deployment Working Group[1] or the

[1] http://www.w3.org/2001/sw/BestPractices/

European research projects OntoWeb and KnowledgeWeb.[2]

In this paper we aim at modestly contributing to these efforts by reporting our practical experiences and lessons learned in building an ontology-based application in the medical domain. The project we present uses domain ontologies and ontology-driven natural language processing to support a content-based retrieval of medical information in textual and image form. In trying to develop the target application ontology we investigated two engineering alternatives: 1). the reuse of the huge amount of domain knowledge already available in ontology-like form in the medical domain, followed by 2). the usage of domain-centered text documents as an input for the domain conceptualization. The second engineering experiment was triggered by the poor fitness of use of the reuse-derived medical ontology in the application context (as described below). Our experiences confirm previous findings in the knowledge acquisition literature and in recent surveys of the state of the art in the ontology engineering area: 1). building ontology-based applications is still a tedious process because of the lack of proved and tested support tools and methods; and 2). reusing existing ontologies in new application contexts is currently related to considerable efforts which might outreach the costs of a new implementation.

The setting of our project is often considered a typical use scenario for the deployment of Semantic Web technologies, both w.r.t the application type (i.e. information retrieval) and the application domain (i.e. medicine)[3]. Given the complexity of the medical domain and the lack of operational experience in this field in industrial context, practice-oriented case studies and guidelines for building Semantic Web medical applications are a core requirement for a serious impact of Semantic Web technologies in the medical domain. Therefore, a second goal of this paper was to use the experiences gained during the project to further elaborate existing best practices[3] towards a list of recommendations for the medical domain, which, far beyond from claiming for completeness, might be useful for developing similar real-world applications.

The rest of the paper is organized as follows: Section 2 gives a brief overview of the application scenario. Section 3 describes the generation of the ontology used by the retrieval system and summarizes lessons learned during this project phase. We conclude with extending existing, domain-independent guidelines for building Semantic Web retrieval applications[3] according to the empirical findings of the project (Section 4).

2 The Project Setting

The aim of the project "A Semantic Web for Pathology" was the realization of a content-based retrieval system for the domain of *lung pathology*.[3] The application

[2] http://ontoweb.org, http://knowledgeweb.semanticweb.org/index.html.

[3] This work has been partially supported by the EU NoE "KnowledgeWeb" (FP6-507482). The project "A Semantic Web for Pathology" is funded by the DFG (German Research Foundation) and is a cooperation between our working group, the Charité Hospital in Berlin, Germany and the Department of Appl. Computer Linguistics at the University of Potsdam, Germany.

uses an explicit semantic representation of the text documents (i.e. pathology reports), which are linked to the semantically annotated digital images they describe[11]. The semantic representation of both documents and images is aligned to a pre-defined ontology, which is used to control the annotation vocabulary and to assist the linguistic analysis of the textual data. Further on, by using background domain knowledge, the system is able to answer content-related queries on image-based data (such as "images where the tissue sample presents certain morphological characteristics"), but also to compare textual medical reports or support differential diagnostics tasks (i.e. retrieve reports with a similar appearance, but alternative diagnosis).

The most important design criterion to obtain a high user acceptance was the seamless integration of the system into the workflow of a pathologist. This minimal invasive usage is the result of a careful design of the user interface, query language and multi-modal content presentation methods. The system offers the ability to write a medical report while examining the images with the conventional microscope or related software, thus restoring the current missing link between text and image-based data.[4] A new patient report is annotated with ontology concepts by a text processing component[9] and the resulting semantic representation is forwarded to a quality assurance module to check its validity w.r.t. the content of the knowledge base. The annotated report is stored in a report repository, while the instantiated concepts and relations extracted during the linguistic analysis (i.e. its semantic representation) become part of the knowledge base. A retrieval component processes several categories of free text queries and returns the suitable images and text documents. In order to enable the user to explore the available expert knowledge, the retrieval component offers multi-modal representations of the query results, which are presented textually and graphically in the context of the domain ontology, and suggests similar documents, query refinements or reformulations. Furthermore, the medical data (e.g. patient reports, digital images, as well as ontological knowledge) is shared or exchanged among domain experts and health-care organizations. Therefore the system makes use of standard, platform independent technologies which support these requirements, like Semantic Web technologies, established XML-based formalizations of patient records such as HL7, and Web Services. It has a Web-based distributed architecture to simplify the data exchange among health-care institutions and insurance companies and to enable a collaborative evolution and deployment of the domain ontology.

The system components and their main features are described in detail in [11]. At present we developed the medical knowledge base (see Section 3) and the client application including a component to generate and edit medical reports and annotate digital images and an ontology-based retrieval component. The client communicates with a number of Web services, which analyze, annotate and store the given information and query the knowledge base. A variety of methods, tools and technologies have been used to create, store, query, inference

[4] Refer to [11] for a detailed description of the application setting w.r.t. "Digital Pathology".

and extend the domain ontology. In this paper we focus solely on the insights gained during the knowledge acquisition phase; practical experiences in using existing tools such as ontology editors, reasoners or APIs are described in [5].

3 Building the Medical Ontology

In the last decades a wide rage of methodologies for building ontologies have been proposed [2]. Apart from minimal differences related to domain and application constraints, ontology engineering methodologies usually introduce an iterative process consisting of the following phases: i). **domain analysis** (including requirements analysis and knowledge acquisition); ii). **conceptualization of the domain knowledge**; iii). **ontology implementation**, iv). **ontology population** (i.e. generation or integration of ontology instances), v). **ontology evaluation**, vi). **ontology refinement**, and vii). **ontology maintenance**. In our setting we identified the following subtasks, which guided the implementation of the ontology for lung pathology:
1. Analysis of the application domain During intensive collaboration with domain experts we specified the thematic clusters and the core features of the ontology. We identified several classes of generic queries which should be covered by the target ontology, while the form of the ontology was dictated by its usage in the language-driven semantic annotation process. In this step potentially relevant knowledge sources were identified: a corpus of patient records (approx. 750 documents), quality guidelines for medical reports used in the health-care organization involved in the project, and medical ontologies such as UMLS, SNOMED, ICD, GeneOntology, to name only a few.
2. Ontology conceptualization In a first experiment the conceptualization step corresponds to the customization of UMLS to the application domain.[5] Due to the suboptimal outcomes of this attempt we decided to perform a second engineering experiment, in which we used the aforementioned corpus of patient records and simple linguistic tools to aid the ontology engineer and in particular the domain expert in specifying ontology concepts and their interdependencies along semantic relationships.
3. Ontology implementation After analyzing the requirements w.r.t. the representation language expressivity, the resulting ontologies were (semi-automatically) implemented in OWL. Quality guidelines describing the content and the structure of the medical reports were exemplarily formalized using SWRL and Jess.
4. Ontology evaluation and refinement A first evaluation of the ontologies was realized though a comparison of the ontology domain coverage with a set of domain-relevant documents, followed by a domain-expert-driven assessment phase.[6] We are currently evaluating the "text-close" ontology w.r.t. its usability in semantic annotation tasks as part of the envisioned medical information system. The first ontology, which was generated by reuse from existing medical

[5] http://www.nlm.nih.gov/research/umls

[6] In the second, corpus-driven ontology engineering experiment the ontology was generated from a "training" set of approximately 400 domain-specific documents, while the remaining 350 documents were used to evaluate the results.

ontologies, was not included to the final application because of its poor fitness of use and user acceptance.

In the following we give some details about the generation of the ontologies and conclude with lessons learned during this attempt. Due to space considerations we focus on reporting on our experiences in the reuse-oriented setting, while the second one is described in [4].

3.1 Reusing Available Medical Knowledge

Due to the difficulties and costs involved in building an ontology from scratch, methodologies often recommend to rely on available domain-related ontologies in order to simplify the domain analysis and the conceptualization phases. However, most of the methodologies deal with this topic only in passing, and there is no common practice for reusing current knowledge sources for generating new ontologies.[7] For example, Uschold and King [13] describe in detail how to build an ontology from scratch, but on the matter of ontology reuse they only give some very general recommendations and explanations of the approach.

Likewise other engineering disciplines reuse implies finding, evaluating, customizing and integrating existing (reusable) components to a target system [7, 12]. The evaluation process requires the specification of a quality framework for ontologies, which addresses both general and application-dependent quality criteria. A complete evaluation of the plethora of medical ontologies currently available was of course unrealistic for our purposes, due to the large number, but also due to the complexity of each of the candidate ontologies. [8] Therefore we employed a restricted quality model, similar to the one recommended in [7], consisting of the following criteria: i). domain coverage: *to what extent does the source ontology cover the medical sub-domains which are relevant for the application ontology (e.g. anatomy and diseases of the lung, genetics, immunohistology)?*, ii). syntax: *is the ontology available in electronic form and what tools are provided to translate it in other representation forms?*, and iii). usage: *to what extent can the ontology be used in the semantic annotation of natural language patient records? To what extent can the ontology be managed by humans in navigating the corpus of patient records?*. The evaluation procedure w.r.t. the first two criteria resulted in a set of approximately 10 medical ontologies, which were without exception integrated in UMLS, containing approximately 250,000 different concepts. The size of the concept set can be explained if we consider the fact that the UMLS knowledge is concentrated in a few major libraries (e.g. MeSH, SNOMED98),[9] which cover important parts of the complete thesaurus. Managing an ontology of such dimensions with Semantic Web technologies is still related to major scalability and performance pitfalls. Besides, building the

[7] By knowledge sources we mean not only ontologies in a broad sense, i.e. taxonomies, thesauri, OWL-ontologies, ER-models and UML class diagrams, but also documents in natural language.

[8] The starting set of candidate ontologies contained over 100 elements. Most of candidate ontologies contained over 5000 concepts.

[9] http://www.nlm.nih.gov/mesh/meshhome.html, http://www.snomed.org

knowledge base implies also a subsequent adaptation and daily usage of its content, a task which is performed by domain experts. Apart from technical drawbacks, very large ontologies can not be used efficiently by humans as well.

Therefore, in order to differentiate among the concepts within the remaining 10 libraries, pathology experts selected 4 central concepts in lung anatomy (i.e. "lung", "pleura", "trachea" and "bronchia") and extracted similar or related concepts from the libraries. They considered the list of all distinct concepts related through a relation of any kind[10] to the 4 initial concepts. The result was a set of approximately 1000 concepts describing the anatomy of the lung and lung diseases and served as initial input for the application ontology.

The linguistic analysis of the patient report corpus evidenced the content-related limitations of UMLS w.r.t. the concrete vocabulary of former. These concepts are really used by pathologists when putting down their observations and will therefore also occur as search parameters when using the retrieval system. We modeled additional, pathology-specific concepts (approximately 200), such as the components and typical content of a medical report, and integrate them in the available ontology library. Besides content-related adaptation needs, the analysis of the generated ontology outlined the absence of concept names in the German language: due to the predominance of English in denominating UMLS concepts and the predominance of German terms in the pathology report archive we had to translate the English terms in order to enable an automatic, ontology-driven annotation of the documents.

After identifying the relevant knowledge sources and the list of concepts which can be used as input for our application, we translated the UMLS data model to the OWL model and transformed the relevant data from one format to another[6, 5]. We implemented a Java-based module, which reads the UMLS data from a relational database and generates the corresponding OWL constructs using Jena2. The resulting ontologies are published server-side and can be accessed by all components of the system.

3.2 Lessons learned

Reusing medical ontologies for the generation of the target application ontology proved to be a tedious, time-consuming and error-prone process. The success of reusing ontologies was influenced to a large extent by two factors: the cost-effective evaluation and customization of the source ontologies[11].

Typically ontology reuse starts with the identification and evaluation of knowledge sources potentially useful for the application domain, which differ both in the covered content, and in the formalization (thesauri, XML-Schemes and DTDs, ER models, UML diagrams, textual descriptions etc.). An automatic integration of the source ontologies means not only the translation of the representation languages to a common format, but eventually also the matching of the resulting schemes. Due to scalability and heterogeneity issues both of

[10] The UMLS Metathesarus contains 7 important relations between concepts: parent, child, sibling, narrower, broader, related-other, source-synonymy.

[11] These tasks required more than 50% of the total engineering time.

these steps could not be performed efficiently using current techniques in our project setting. However our experiences showed that using simple techniques which rely on the smallest common denominator of the employed knowledge sources—usually their vocabulary— significantly increased the efficiency of the reuse process. This finding was confirmed by an analysis of the resulted ontology w.r.t. its dependencies to the source ontologies: the target ontology reuses their vocabulary to a large extent—properties and axioms are not supported explicitly in many knowledge sources and their integration to the target ontology was non-trivial for both humans and tools.

UMLS contains several problematic modeling decisions which have been often described in research projects aiming to integrate it in knowledge-based applications[10, 1, 8]. In our application setting the UMLS-based ontology showed important limitations especially w.r.t. its usability for semantic annotation. The absence of concept names in a linguistically predictable form decreased the quality of the automatic annotation drastically. An optimal usage of UMLS in similar settings will therefore require a comprehensive analysis of the denomination types across the UMLS libraries.

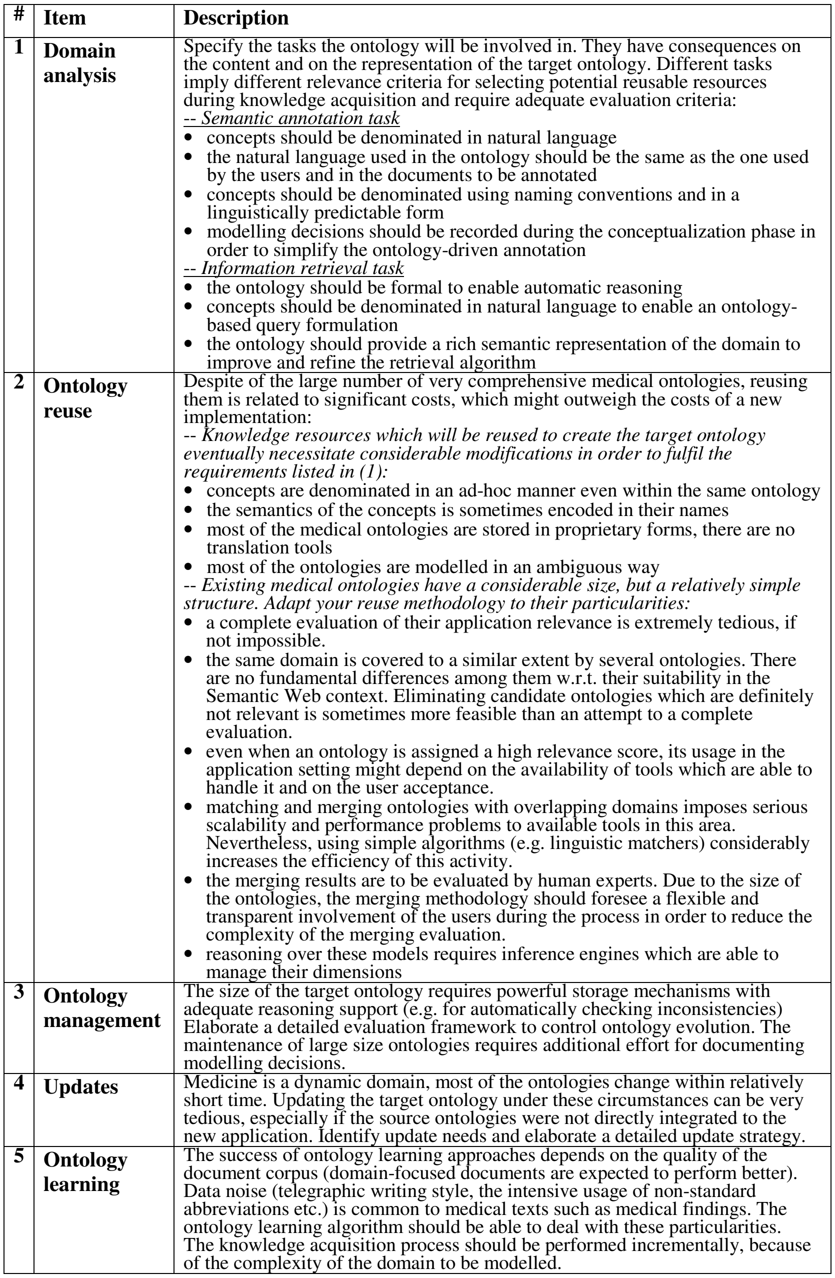
Another important issue was the usage of the ontology by the community of domain experts, which reported serious acceptance problems w.r.t. the UMLS-based ontology: domain experts seemed to have difficulties in trusting the content of the ontology and in methodically extending it for a more detailed representation of pathology-specific knowledge. This was the main motivation for alternatively building a second application ontology on the basis of the domain corpus of patient records provided by our health-care partner[4].

The engineering process relied on the same engineering methodology as the first experiment, while XML-based medical reports were employed as an input for the conceptualization phase. [12] The main advantages of the latter experiment compared to the UMLS-based one were the significant cost savings and the increased fitness of use of the generated ontology w.r.t. the semantic annotation task. From a resource point of view, building the first ontology involved four times as many resources than the second approach (5 person-months for the UMLS-based ontology with 1200 concepts vs. 1.25 person-months for the "text-close" ontology of a similar size). The evaluation of the suitability of the two ontologies to semantically annotating medical documents (described in [4]) confirmed the results of the resource-based evaluation. Orthogonal to technical and economical benefits the ontology derived from the medical reports had a considerably higher acceptance rate among its users: the results of the methodology were easily understandable to the domain experts, who were rapidly able to evaluate and refine the ontology.

4 Preliminary guidelines

According to our findings in the mentioned setting we conclude this paper with a (incomplete) list of guidelines for building Semantic Web retrieval applications in

[12] Refer to [4] for a detailed description and evaluation of the methods employed.



#	Item	Description
1	**Domain analysis**	Specify the tasks the ontology will be involved in. They have consequences on the content and on the representation of the target ontology. Different tasks imply different relevance criteria for selecting potential reusable resources during knowledge acquisition and require adequate evaluation criteria: *-- Semantic annotation task* • concepts should be denominated in natural language • the natural language used in the ontology should be the same as the one used by the users and in the documents to be annotated • concepts should be denominated using naming conventions and in a linguistically predictable form • modelling decisions should be recorded during the conceptualization phase in order to simplify the ontology-driven annotation *-- Information retrieval task* • the ontology should be formal to enable automatic reasoning • concepts should be denominated in natural language to enable an ontology-based query formulation • the ontology should provide a rich semantic representation of the domain to improve and refine the retrieval algorithm
2	**Ontology reuse**	Despite of the large number of very comprehensive medical ontologies, reusing them is related to significant costs, which might outweigh the costs of a new implementation: -- *Knowledge resources which will be reused to create the target ontology eventually necessitate considerable modifications in order to fulfil the requirements listed in (1):* • concepts are denominated in an ad-hoc manner even within the same ontology • the semantics of the concepts is sometimes encoded in their names • most of the medical ontologies are stored in proprietary forms, there are no translation tools • most of the ontologies are modelled in an ambiguous way -- *Existing medical ontologies have a considerable size, but a relatively simple structure. Adapt your reuse methodology to their particularities:* • a complete evaluation of their application relevance is extremely tedious, if not impossible. • the same domain is covered to a similar extent by several ontologies. There are no fundamental differences among them w.r.t. their suitability in the Semantic Web context. Eliminating candidate ontologies which are definitely not relevant is sometimes more feasible than an attempt to a complete evaluation. • even when an ontology is assigned a high relevance score, its usage in the application setting might depend on the availability of tools which are able to handle it and on the user acceptance. • matching and merging ontologies with overlapping domains imposes serious scalability and performance problems to available tools in this area. Nevertheless, using simple algorithms (e.g. linguistic matchers) considerably increases the efficiency of this activity. • the merging results are to be evaluated by human experts. Due to the size of the ontologies, the merging methodology should foresee a flexible and transparent involvement of the users during the process in order to reduce the complexity of the merging evaluation. • reasoning over these models requires inference engines which are able to manage their dimensions
3	**Ontology management**	The size of the target ontology requires powerful storage mechanisms with adequate reasoning support (e.g. for automatically checking inconsistencies) Elaborate a detailed evaluation framework to control ontology evolution. The maintenance of large size ontologies requires additional effort for documenting modelling decisions.
4	**Updates**	Medicine is a dynamic domain, most of the ontologies change within relatively short time. Updating the target ontology under these circumstances can be very tedious, especially if the source ontologies were not directly integrated to the new application. Identify update needs and elaborate a detailed update strategy.
5	**Ontology learning**	The success of ontology learning approaches depends on the quality of the document corpus (domain-focused documents are expected to perform better). Data noise (telegraphic writing style, the intensive usage of non-standard abbreviations etc.) is common to medical texts such as medical findings. The ontology learning algorithm should be able to deal with these particularities. The knowledge acquisition process should be performed incrementally, because of the complexity of the domain to be modelled.

Fig. 1. Guidelines for Building Ontology-based Retrieval Applications

the domain of medicine. As a starting point we used a set of domain-independent guidelines emerged in the European project OntoWeb, which focus less on technical aspects, but mainly on "issues that relate to the business environment that affects the deployment, integration and acceptance of the ontology-based application"[3]. The initial checklist contains 13 items, which cover both organizational and ontology-specific issues. Since we did not encounter any problems related to the organizational setting (satisfactory user involvement, no legacy systems or licence problems etc.), we elaborated the topics which relate directly to the ontology engineering process and adapted them to the medical domain. The results of our analysis are summarized in the table above (Figure 1). This list could be complimented with modeling guidelines, which are equally important in a complex domain such as medicine. Such guidelines are currently emerging as a result of the initiatives of the W3C Semantic Web Best Practices and Deployment Working Group.

References

1. A. Burgun and O. Bodenreider. Mapping the UMLS Semantic Network into General Ontologies. In *Proc. of the AMIA Symposium*, 2001.
2. M. Fernández-López and A. Gómez-Pérez. Overview and Analysis of Methodologies for Building Ontologies. *Knowledge Engineering Review*, 17(2):129–156, 2002.
3. OntoWeb European Project. Successful scenarios for ontology-based applications (Deliverable D2.2 OntoWeb IST-2001-29243), 2002.
4. E. Paslaru Bontas, D. Schlangen, and T. Schrader. Creating ontologies for content representation — the OntoSeed suite. In *Proc. of the Int. Conference ODBASE2005 (to be published)*, 2005.
5. E. Paslaru Bontas, S. Tietz, and T. Schrader. Experiences Using Semantic Web Technologies to Realize an Information Retrieval System for Pathology. In *Proc. of the Berliner XML Days Conference*, 2004.
6. E. Paslaru Bontas, S. Tietz, R. Tolksdorf, and T. Schrader. Generation and Management of a Medical Ontology in a Semantic Web Retrieval System. In *Proc. of the CoopIS/DOA/ODBASE Conferences*, 2004.
7. H. S. Pinto and J. P. Martins. A methodology for ontology integration. In *Proc. of the Int. Conference on Knowledge Capture K-CAP01*, 2001.
8. D.M. Pisanelli, A. Gangemi, and G. Steve. Ontological Analysis of the UMLS Metathesaurus. *JAMIA*, 5:810 – 814, 1998.
9. D. Schlangen, M. Stede, and E. Paslaru Bontas. Feeding OWL: Extracting and Representing the Content of Pathology Reports. In *Proc. of the NLPXML 2004*, 2004.
10. S. Schulze-Kremer, B. Smith, and A. Kumar. Revising the UMLS Semantic Network. In *Proc. of the Medinfo 2004*, 2004.
11. R. Tolksdorf and E. Paslaru Bontas. Organizing Knowledge in a Semantic Web for Pathology. In *Proc. of the NetObjectDays Conference 2004*, 2004.
12. M. Uschold, M. Healy, K. Williamson, P. Clark, and S. Woods. Ontology Reuse and Application. In *Proc. of the Int. Conf. on Formal Ontology and Information Systems FOIS98*, 1998.
13. M. Uschold and M. King. Towards a Methodology for Building Ontologies. In *Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI'95*, 1995.