Ontology-based Knowledge Organization in a Semantic Web for Pathology

Elena Paslaru Bontas[1], Robert Tolksdorf[1], Thomas Schrader[2]

[1]Freie Universität Berlin
Institut für Informatik
AG Netzbasierte Informationssysteme
Takustr. 9, D-14195 Berlin, Germany
paslaru@inf.fu-berlin.de, *research@robert-tolksdorf.de*
[2]Institute of Pathology Charité
Rudolf-Virchow-Haus
Schumannstr. 20-21
D-10117 Berlin, Germany
thomas.schrader@charite.de

Abstract: Digital pathology can be defined as the realization of a completely digitized diagnostic process in pathology with fully scanned glass slides. As a consequence, a competitive digital pathology tool should be able to manage the resulting digitized data and allow the pathologists to retrieve and visualize medical reports and corresponding digital slides efficiently. In this paper we present a content-based retrieval system for text and image data for the domain of "lung pathology". The system is intended to overcome the limitations of current approaches in digital pathology, by using a medical knowledge component formalized with Semantic Web representation languages and a semantically-annotated pathology data archive.

Key Words: Semantic Web, Knowledge Management, Telepathology, Digital Pathology

1 Digital Pathology and Telepathology

Pathology is a theoretical subject of medical science and plays an important role for the diagnosis of diseases in clinical medicine. Apart from the autopsy diagnosis, the morphological (histological) examination of tissue with a microscope is the main task in pathology.

To make a diagnosis the physician usually takes a tissue sample. This material is sent to the Institute of Pathology and after a preparation procedure the pathologist uses glass slides to examine the tissue using a conventional microscope (Figure 1). A pathology report describes the abnormality and concludes with a diagnosis which is used as the basis for the further therapeutic process. In most cases a histological, morphological diagnosis is a pre-condition for the therapeutical decision.

The pathology report consists of three main parts: macroscopy (a description of the tissue sample), microscopy (a histological description of the abnormality) and the diagnosis (the conclusion about the clinical data, patient's age and gender, macroscopy and microscopy). Sometimes a commentary is added to explain the diagnosis and the consequences as well as a tumor classification code.

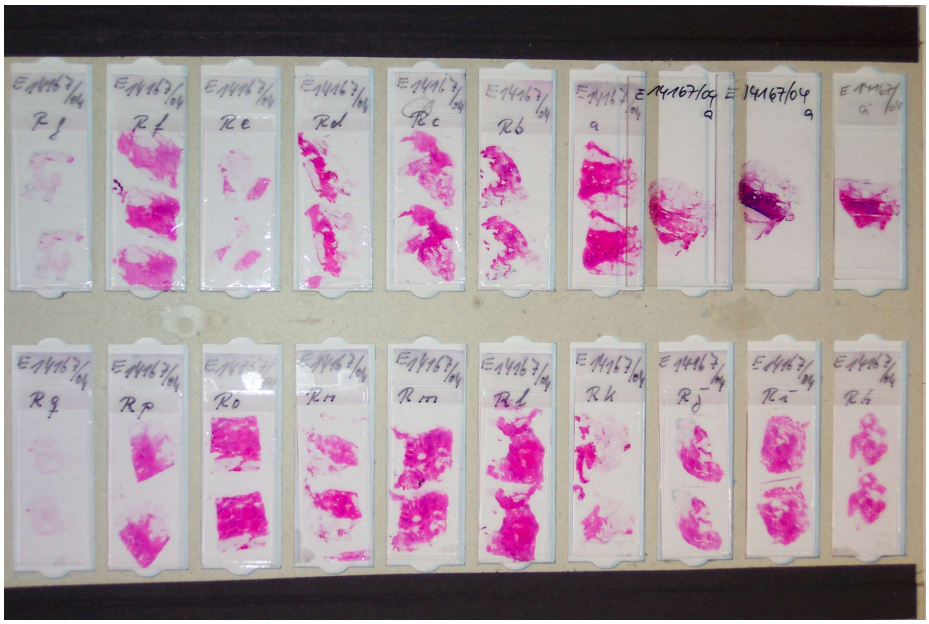

Figure 1: Conventional Glass Slides for Diagnostic Pathology

In most cases a fully accredited pathologist can diagnose the abnormalities. In 5% to 20% of cases it is necessary to consult an expert in a special topic (second opinion) because the tumor or abnormality is rare or complicated. Normally such cases are sent to an expert by mail. With the development of the Internet and the improvement of telecommunication telepathology can ease the acquisition of a second opinion. Some digitized images of the abnormality are sent to a telepathology consultation center such the UICC-TPCC. The case with the clinical data and the images can be viewed by an internationally accepted expert and the requesting pathologist receives advice or a diagnosis [7].

The latest technological advances in pathology have been influenced by telepathology, i.e. the possibility to exchange digitized images with an expert for a second opinion. Remotely controlled microscopes meet the requirements of telepathology through the use of an attached video camera and a motorized cross table. Tools for digital pathology are now coming on the market, for example slide scanners which use new technologies to store, transfer and load large images efficiently. Summing up, digital pathology can be defined as the realization of a completely digitized diagnostic process in pathology with fully scanned glass slides.

The Institute of Pathology, Charité, has developed the Digital Virtual Microscope (DVM) to load and present large images of about 5 GB in real time without any image

quality loss [5] (Figure 2). The routine use of a virtual microscope will be a paradigm shift comparable with the introduction of the digital radiology. For the first time all histological glass slides will be scanned and digitized and can be viewed by a specialist browser in real time. For example, in a large-sized institute such a tool has to store and process more then 200 cases with approximately 2000 images on a daily basis. The main advantages of a virtual microscope are the concurrent, time independent and remote usage of all digitized images and their reuse for other purposes.

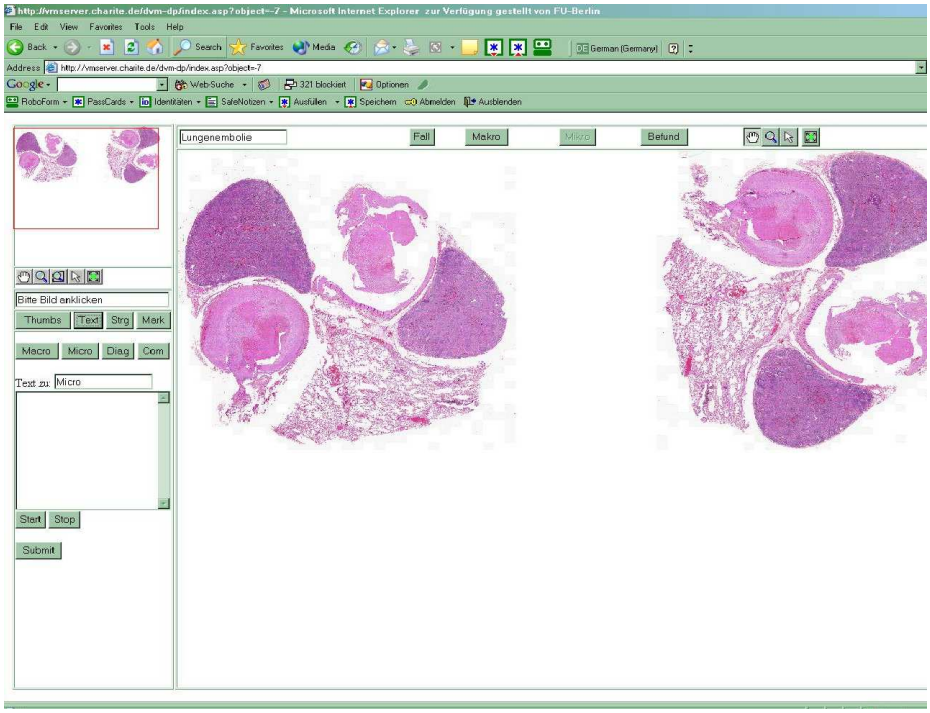

Figure 2: The Digital Virtual Microscope

2 Information System Support for Pathology

As a consequence of this new procedure of digitizing all medical cases, pathology systems must be able to handle very large image databases. At Charité Berlin we already utilize a Pathology Laboratory Information System (PLIS) as a document management system. The main limitation of such a system is the missing link between pathology

reports and images and especially the content of images. To make use of a case for a presentation or for educational purposes the pathologist has to create digital images with a microscope camera. The resulting files are stored on the personal computer ordered by case number or by diagnosis. Other pathologists can not use these images since they can not be retrieved or searched over a network. The usage of an image database would only be a partial solution of the retrieval problem because it would not create the link between a diagnosis and the content of an image. Since a digitized image is very huge with a file size between 1 and 15 GB (printed it would occupy a space of 70 m^2) the connection between diagnosis/report and corresponding image can be used to retrieve and visualize only relevant image fragments instead of complete images.

Through observation of the typical work practise of the pathologist we developed the model of the ”diagnostic path” [8]. The point of time in which the pathologist dictates this chapter correlates of course with his current position in the slide. Hence, in the DVM we store the movement path through the slide (the so-called ”observation path”) and match this with the time points of the corresponding textual observations (the so-called ”dictation path”). This diagnostic path is the missing link between the content of an image and the pathology report.

Besides considering the connection between textual reports and images, an efficient retrieval tool should also support various use cases in pathology:

- the diagnostic process by presentation of cases in relation to the problem (same diagnosis or differential diagnosis),
- quality management by analysis of the reports,
- the reuse of cases and their images - the digital slides - for further purposes in research and education.

3 A Semantic Web for Pathology

In order to achieve the goals outlined in Section 2 we need to improve the retrieval capabilities of the pathology data archive (i.e. pathology reports and digital histological slides). For this purpose the system should offer the pathologists intuitive information access, which implies minimal technical know-how w.r.t. the underlying storage system and its query language. Besides, it should go beyond standard database search techniques and allow a content-based retrieval of the (textual) pathology reports and the associated digital images. By using the pathology reports as semantic metadata of the images the system supports knowledge-intensive query capabilities beyond string matching and image-based algorithms. Queries like “images where no tumor has been reported”, “similar cases” or “images containing a tumor of 15mm in diameter” can be answered only in a setting where text and image-based information is integrated and annotated with explicitly represented semantic connotations.

The project “A Semantic Web for Pathology” aims to put these ideas and requirements into practice by realizing a Semantic-Web based retrieval system for the domain of ”lung pathology”. For this purpose the pathology data is annotated with semantic references and the textual pathology reports are used as descriptions of what the associated images represent. The annotation process is realized using ontology-based text-processing algorithms i.e. NLP heuristics using a domain ontology representing background medical knowledge, which also serves as a vocabulary in order to control the retrieval process. A detailed description of the system and its components is presented in [12].

3.1 Re-structuring the Pathology Archive

The first step for the realization of the system was the re-organization of the available information. We took a subset of approximately 700 lung pathology reports from the Institute of Pathology, Charité and some of the corresponding digital images. For this purpose we defined an XML-schema based on the HL7/CDA medical standard (reference) which includes application-relevant data of the patient record (like age and gender) and reflects the typical structure of a pathology report (see Section 1).

The connection to the digital images is fundamental for efficient retrieval. It should return both the relevant textual information and the corresponding image region the pathologist refers to in a certain portion of text. Since the size of such images is up to 15 GB, it is more useful to retrieve a concrete image sector than complete images for a certain user query. For this purpose we use the functionality of the DVM (see Section 1). The virtual microscope stores information related to the “diagnostic path” (see Section 2) which mirrors the way the diagnosis decision was accomplished. A fragment of a XML-encoded medical report (in German) is presented in Figure 3.

Besides knowledge relevant for pathology, which is grouped in the content of the tags *macroscopy*, *microscopy*, *diagnosis* and *comments*, every XML pathology report contains elements required by the HL7/CDA standards: administrative and actuarial information, e.g. legally responsible pathologist and medical organization. Note also the *coded_entry* elements containing the reference to the corresponding image and time stamps of the dictation path (see Section 2). The standards requirements influence the XML format of the reports in a sometimes counter-intuitive way (the information must be encoded within *section*, *paragraphs* and *coded_entry* tags), but they are a prerequisite for the exchange and reuse of information in networked systems.

On the basis of the designed XML-schema, we transferred the available pathology reports into the new XML format and implemented an Java-based report editor which can be integrated in the current functionality of the DVM (see Section 2). This results in a convenient tool which allows pathologists to observe high-quality digital histological slides, annotate them and in the same time edit or view the associated medical report. Information related to the dictation and observation paths is also recorded for every user operation (e.g. text editing, zoom in).

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>
<levelone
 xmlns="urn::hl7-org/cda" xmlns:sciphox="urn::sciphox-org/sciphox"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="urn::hl7-org/cda sciphox-cda.xsd"
 xmlns:swpatho="urn::swpatho-org">
 <clinical_document_header>
  <id EX="149331"/>
  <document_type_cd V="11529-5" S="2.16.840.1.113883.6.1"
  DN="Pathologischer Bericht"/>
  <origination_dttm V="19.2.2001 00:00:00"/>
  <patient_encounter><id EX="129313"/>
   <practice_setting_cd V="HOSP" S="2.16.840.1.113883.5.10588"
   DN="Pathologie"/>
   <encounter_tmr V="16.2.2001 00:00:00" />
  </patient_encounter>
  <authenticator>...</authenticator>
  <legal_authenticator>...</legal_authenticator>
  <intended_recipient>...</intended_recipient>
  <originator><originator.type_cd V="AUT"/>
   <participation_tmr V="19.2.2001 00:00:00"/>
   ...</originator>
  <originating_organization>...
   <organization><organization.nm V="Institut für Pathologie,Charite"/>
   </organization>
  </originating_organization>
  <provider>...</provider>
  <patient>
   <patient.type_cd V="PATSBJ"/>
   <person><id EX="44986"/>...</person><birth_dttm V="64"/>
   <administrative_gender_cd V="M" />
   <local_header ignore="all" descriptor="swpatho">
    <swpatho:swpatho-ssu type="Kostentraeger" country="de" version="v1">
    <swpatho:Kostentraegerbezeichnung V="CHA"/></swpatho:swpatho-ssu>
    <swpatho:swpatho-ssu type="Schreibkraft" country="de" version="v1">
    <swpatho:Schreibkraftkuerzel V="SGGX"/></swpatho:swpatho-ssu>
    <swpatho:swpatho-ssu type="E-Nummer" country="de" version="v1">
    <swpatho:E-Nummer V="E06823-01"/></swpatho:swpatho-ssu>
   </local_header>
  </patient>
 </clinical_document_header>
 <body>
  <section><caption>Befund</caption>
  <section><caption>Makroskopie</caption>
   <paragraph><content>Sechs PE.</content></paragraph>
  </section>
  <section><caption>Mikroskopie</caption>
  <coded_entry><coded_entry.value V="5" S="UID" /></coded_entry>
  <coded_entry><coded_entry.value V="6" S="PID" /></coded_entry>
  <coded_entry><coded_entry.value V="Feb 09 13:53:16 CET 2004"
  S="StartTime"/></coded_entry>
  <coded_entry><coded_entry.value V="Feb 09 13:53:18 CET 2004"
  S="StopTime"/></coded_entry>
  <paragraph><content>Mehrere respiratorische Schleimhautlamellen sowie
   Lungenparenchymanteile, das vorhandene Epithel einreihig und
   dysplasiefrei, das Stroma entzündlich infiltriert, mit
   reichlich epotheloidzelligen Granulomen ohne Nekrose, Nachweis
   von zahlreichen mehrkernigen Riesenzellen.</content></paragraph>
  </section>
  <section><caption>Kritischer_Bericht</caption>
   <paragraph><content>Respiratorische Schleimhautlamellen sowie
   Lungenparenchym mit nichtverkäsender epitheloidzelliger granulomatöser
   Entzündung.</content></paragraph>
  </section>
  <section><caption>Kommentar</caption>
   <paragraph><content>Der histologische Befund entspricht einer Sarcoidose.
   Kein Hinweis für Malignität oder PCP.</content></paragraph>
  </section>
  </section>
 </body>
</levelone>
```

Figure 3: Fragment of a medical report in XML

3.2 Semantic Annotation of the Pathology Archive

The transformation of reports in a semi-structured format (e.g. the XML-based format CDA/HL7) does not make any content related information available. To achieve a significant improvement in retrieval, one needs an explicit representation of the semantics of the content. We build an ontology-based component to recognize concept instances referencing a domain ontology. The component makes use of ontology-driven text processing algorithms. It parses the text of the pathology reports and extracts potentially relevant concepts (nouns, nominal phrases), which are mapped to ontology concepts from the knowledge base. The detailed realization of the annotation process is presented in [6].

3.3 The Pathology Knowledge Base

At the core of the retrieval system is a knowledge base, which is used both for the semantic annotation of the pathology data (Section 3.2) and for content-based retrieval and diagnosis quality management (e.g. proving if the facts within a medical report are contradictory w.r.t. the background knowledge of the system). It contains an ontology library and a rule set and is implemented with Semantic Web technologies [3]. The usage of Semantic Web technologies (representation languages, domain ontologies) supports the sharing and exchange of the pathology data among domain experts and health-care organizations, which is one of the main requirements when realizing a Digital Pathology tool (see Section 1.

The domain ontology contains background information from the application domain "lung pathology". As input for the domain ontology we used UMLS [13], which is at present the most comprehensive medical thesaurus, integrating more than 100 medical libraries in a common data format. Due to its dimensions and broad thematic coverage of the libraries, the integration of the complete thesaurus in the application knowledge base is not feasible. Neither human experts nor available Semantic Web technologies are able to manage and evaluate it efficiently.

For the realization of a domain ontology customized for the concrete application scenario we selected a set of approximately 1000 concepts related to central concepts in lung pathology e.g. "lung", "bronchia", "trachea" and "pleura". After their selection in collaboration with domain experts, they were represented with the Web Ontology Language OWL [2]. The domain knowledge was extended by modeling additional pathology-specific knowledge, not present in UMLS. We focused on concepts which are frequently used in the pathology report archive. These concepts are part of a lexicon generated during the lexical analysis of the archive and can be divided in the following thematic categories:

– *Medical concepts* concerning the anatomy of lung or lung diseases, which are not contained in UMLS or have not been selected as being relevant by our pre-selection

step. Another issue related to the content limitation of UMLS and implicitly of our core domain ontology is the poor coverage of terms in the German language (from over 500.000 concepts, approximately 12.000 concepts were given in German in the 2003AC UMLS release).

- *Typical structure of the pathology report* (see Section 2), which also mirrors the typical subtasks of the diagnosis procedure.
- *Typical pathology concepts*, i.e. concepts closely related to routine pathology work, which occur in most of the pathology texts.
- *Non-medical concepts*, which are very frequent in the macroscopy and microscopy fragments of pathology reports (see Section 3.1). They usually describe spatial relationships and properties of physical concepts (e.g. length, color, texture, form, diameter etc.). We do not aim at a very detailed and complete representation of non-medical terms. Rather we will incrementally extend their modeling dependent on the quality of the retrieval results.

A second component of the knowledge base consists of rules. They are intended to describe on the one hand the decision processes and quality assurance criteria used in pathology and on the other hand to represent domain-specific and even non-medical facts, which can not be expressed using the OWL representation language. The rules will be formalized in RuleML [1] or another related language for rule representation and will be added incrementally depending on the needs of the retrieval services. A core set of medicine-specific rules has already been identified by the domain experts and their integration into the knowledge base is the subject of current work.

4 Changes in Medical Knowledge

In this system we work with medical knowledge. The basis for our ontology is the UMLS thesaurus which collects medical knowledge from a variety of sources. As described in 3.3 we selected a subset of concepts from this, extended it with further concepts and did some corrections of detected errors. In addition, we will add explicit rules to the knowledge base of our system.

The medical knowledge, its representation and the rules are dynamic and change slowly but continuously. This raises several challenges for our system that we outline below.

- *Medical knowledge changes.* This is reflected in new versions of ontologies and rules that are represented in our system. As a consequence of such changes, the existing pathology reports might be inconsistent with the new medical knowledge – with the result that new knowledge cases might be diagnosed differently.

  In our system, we would have to update the OWL representation and the rules to be consistent with the medical knowledge. Given that, we have two cases:

1. We want our system to contain only information that is consistent with the medical knowledge. This is necessary to make similar cases comparable.

   This means however that we would have to restart *the whole process* starting with the ontology-based NLP analysis of reports and their classification. And we would have to revise current results from the inference engine. Making the system consistent with the current medical knowledge seems to be a complex task with unacceptable high costs.

2. We accept that the information stored is not consistent with the current medical knowledge. This seems to be a valid approach, since it reflects the inconsistent state of reports in the real world: Reports are not changed later in the case of new knowledge.

   For our system, however, this implies that any classification of a report is relative to the respective state of the knowledge representation. Two cases that might seem comparable might not be comparable at all when they were analyzed in different knowledge contexts. Over a long time, there will be a substantial *drift* in how documents are classified and compared that leads to an unacceptable imprecision of our retrieval.

   One solution might be to represent the mentioned context explicitly. This would mean to keep versions of the ontologies and rules and to keep track for which version a document has been processed. Explicit descriptions of the changes would have to be taken into account when comparing cases. Again, this approach seems to introduce a high overhead and might even be impossible (see next item).

– *Adjustment of the representation of medical knowledge*. The UMLS thesaurus is the starting point for building our OWL ontologies. A small subset of concepts from the UMLS thesaurus were selected for our system and transferred to OWL. Also, some corrections and extensions were made.

  For any new UMLS version we would have to realign our OWL representation with it. This seems to be non-trivial, since it would require direct intervention of the domain experts.

  Every concept added to the UMLS thesaurus would have to be considered for inclusion into our subset and possibly the consequent inclusion of further concepts and relations. This can only be done manually.

  For every removed concept from the UMLS thesaurus, one needs to determine whether it was removed because it was obsolete, because it is replaced by other new concepts, because its function is better described with other existing concepts etc. Again this might also affect other concepts via relations and again this test has to be performed manually.

In addition, we have made changes in our OWL version of the subset at least by adding information, perhaps adding concepts and perhaps replacing relations, in part motivated by errors in the UMLS [4, 11, 10, 9].

Manual checking again implies an overhead since it has to be done by an expert. We expect that this process will not scale with the number of concepts and the frequency of changes.

We currently see the relation of the UMLS thesaurus to our representation as static and to not implement a change detection in our system. We are exploring to what extent the process can be automated.

5 Conclusion

In this paper we presented our experiences in realizing a Semantic Web-based retrieval system for pathology data. The system is based on a medical knowledge base, which is used to semantically annotate and process available data and user queries. Though the system is still under development, we believe that domain ontologies are a valuable tool to improve the retrieval capabilities of medical information sources. The generation of application-relevant ontologies, which can be embedded efficiently into real-world applications, has proved to be a very challenging task, mainly due to the complexity of the involved knowledge and the lack of reusable knowledge sources. Though Semantic Web technologies still need to demonstrate their feasibility in real-world applications we think at this point that they are a good choice in settings where the reuse, exchange and the shared understanding of knowledge is a key issue.

References

1. The Rule Markup Language (RuleML). http://www.ruleml.org.
2. W3C OWL Reference. http://www.w3.org/TR/owl-ref/.
3. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 5 2001.
4. A. Gangemi, D. M. Pisanelli, and G. Steve. An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies. *Data Knowledge Engineering*, 31(2):183–220, 1999.
5. K. Saeger, K. Schlüns, T. Schrader, and P. Hufnagl. The Virtual Microscope for Routine Pathology based on a PACS system for 6 GB images. In *Proceedings of the 17th International Congress and Exhibition CARS 2003*, 2003.
6. D. Schlangen, M. Stede, and E. Paslaru Bontas. Feeding OWL: Extracting and Representing the Content of Pathology Reports. In *to appear in Proc. NLPXML 2004*, 2004.
7. T Schrader, T. Feig, P. Hufnagl, K. Kayser, and M. Dietel. A userfriendly Telepathology Service at the Internet - The Telepathology Consultation Center of the UICC. *Electronic Journal of Pathology and Histology*, 9(1), 2003.
8. T. Schrader, S. Niepage, T. Leuthold, K. Saeger, S. Hellmig, and P. Hufnagl. Implementation of a pathology report compiler with integrated diagnostic path functionality in the Diagnostic Virtual Microscopy. *Pathol Res Pract*, 200(4), 2004.

9. S. Schulz and U. Hahn. Medical knowledge reegineering - converting major portions of the UMLS into a terminological knowledge base. *International Journal of Medical Informatics*, 2001.
10. S. Schulz, M. Romacker, and U. Hahn. Knowledge engineering the UMLS. *Stud Health Technol Inform*, 77:701–5, 2000.
11. S. Schulze-Kremer, B. Smith, and A. Kumar. Revising the UMLS Semantic Network. In *Proc. Medinfo 2004*, 2004.
12. R. Tolksdorf and E. Paslaru Bontas. Organizing Knowledge in a Semantic Web for Pathology. In *Proc. NetObjectDays 2004*, 2004.
13. Unified Medical Language System. http://www.nlm.nih.gov/research/umls, 2002.