Organizing Knowledge in a Semantic Web for Pathology

Robert Tolksdorf[1], Elena Paslaru Bontas[2]
[1] research@robert-tolksdorf.de, http://www.robert-tolksdorf.de
[2]paslaru@inf.fu-berlin.de

Freie Universität Berlin
Institut für Informatik
AG Netzbasierte Informationssysteme
Takustr. 9, D-14195 Berlin Germany

Abstract. Digital pathology and telepathology allow an extended usage of electronic images for diagnostics, support or educational purposes in anatomical or clinical pathology. Current approaches have not found wide acceptance in routine pathology, mainly due to limitations of image retrieval. In this paper we propose a semantic retrieval system for the pathology domain. The system combines text and image information and offers advanced content-based retrieval services for diagnosis, differential diagnosis and teaching tasks. At the core of the system is a Semantic Web gathering both ontological domain knowledge, and rules describing key tasks and processes in pathology.

1 Introduction

Digital pathology or telepathology intends to extend the usage of electronic images for diagnostical support or educational purposes in anatomical or clinical pathology. The advantages of these approaches are generally accepted and several applications are already available. Nevertheless, none of the available applications has found wide acceptance for diagnostic tasks, mainly due to the huge amount of data resulting from the digitalization process and the limitations of image-based retrieval. In this paper we propose a *semantic* retrieval system for the pathology domain. The system brings both text and image information together and offers advanced content-based retrieval services for diagnosis, differential diagnosis and teaching tasks. At the core of the system there is a Semantic Web gathering both ontological domain knowledge, and rules describing key tasks and processes in pathology. The usage of *Semantic Web* standards and domain ontologies facilitates the realization of a distributed infrastructure for knowledge share and exchange.

The rest of this paper is organized as follows: The remaining introductory sections present the setting of the project, telepathology, and its main ideas and features. Chapter 3 provides an insight into the technical aspects of the retrieval system, by enumerating the technical requirements and the associated system architecture, followed by a detailed description of the system components. At this point we will present our achievements and the challenges we are currently confronted with in the realization of the main components. Chapter 4 delimits our approach from related research efforts in this domain, while Chapter 5 is dedicated to future work.

1.1 Telepathology

Telepathology is a key domain in telemedicine. By using telepathology approaches like virtual microscopy, pathologists analyze high quality digital images on a display screen instead of conventional glass slide at the common light microscope. Such digital images are taken by a camera attached to the microscope and stored for retrieval and reuse (with or without textual annotations) in a database or directly in a patient record.

Health care information systems, which store and integrate information and coordinate actions among health care professionals, have been realized at various places in the last decades. New developments in telemedicine allow medical personnel to remotely deliver health care to the patient. At the Charité Institute of Pathology in Berlin, the first web-based virtual microscope allows histological information to be evaluated, transfered, and stored in digital format [20, 15]. This technique offers essential advantages compared to the classical approach, by supporting communication and exchange among professionals not sharing the same workplace location and improving quality assurance mechanisms [16]. However, to realize a complete computer-based infrastructure for pathology, one needs not only advanced support in the management of digital images. Necessary is also a more efficient integration of the medical reports, which are produced by pathologists to describe their observations from analyzing the slides at the light/digital microscope.

Common information systems in pathology restrict their retrieval capabilities to automatical picture analysis and ignore corresponding medical reports. Such analysis algorithms have the essential drawback that they operate exclusively on structural – or syntactical – parameters such as color, texture and basic geometrical forms while ignoring the real content and the actual meaning of the pictures. Medical reports, however, contain much more than that since they are textual representations of the pictural represented *content* of the slides. By that they capture *implicitly* the actual semantics of what the picture graphically represent, for example “a tumor” in contrast to “a red blob” or “a colocated set of red pixels”. In the project described in this paper, we take the semantics aspects a step further: We understand the reports as semantic metadata for the image prepared by an expert with high quality. We use ontology-based text processing algorithms to make the semantic content *explicit* and build a system that takes advantage of the explicitly represented knowledge.

2 A Semantic Web for Pathology

The project “Semantic Web for Pathology” aims to realize a Semantic Web-based text and picture retrieval system for lung pathology. For this purpose we concentrate our efforts in three interrelated directions: 1) the construction of a *knowledge base*, 2) the development of *knowledge reuse algorithms* and of a 3) *semantic annotation schema* for medical reports and digital histological images.

The knowledge base contains domain ontologies, generic ontologies and rules. Domain ontologies are used for the machine-processable representation of pathology knowledge, while generic ontologies capture common sense knowledge that can be useful in knowledge-intensive tasks. Several very complex libraries of ontologies are already available for this purpose. While ontologies model the background knowledge of the pathologists, the rules are used to describe the decision processes using this knowledge: diagnostics, microscope analysis, observations etc. The acquisition of such rules, which play a crucial role for the retrieval, will be accomplished during an intensive collaboration with domain experts. Further on, we analyze the textual data with text processing algorithms and annotate it with concepts from the knowledge base in order to improve precision and recall in retrieval operations. The annotation scheme is harmonized with the pathology knowledge base by using the corresponding medical ontologies as controlled vocabulary for the annotations. Text analysis is also used to extract implicit factual knowledge, subsequently integrated in the knowledge base.

2.1 Main features

We foresee several valuable uses of the planned system in routine pathology. First, it may be used as an assistant tool for diagnosis tasks. Since knowledge is made explicit, it supports new query capabilities for diagnosis tasks: similarity or identity of cases based on semantic rules and medical ontologies, differential diagnosis, semantically precise statistical information about occurrences of certain distinguishing criteria in a diagnosis case. The provided information will be very valuable in diagnosis work especially for the under-diagnosed cases, since such situations require deeper investigations of the problem domain and a very strict control mechanism of the diagnosis quality ([5]).

Second, advanced retrieval capabilities may be used for educational purposes by teaching personnel and students. Currently, enormous amounts of knowledge are lost by being stored in data bases, which are behaving as real data sinks. They can and should be used for teaching, eg. for case-based medical education.

Third, quality assurance and checking of diagnosis decisions can be effectuated more efficiently because the system uses axioms and rules to automatically check consistency and validity. During the developement phase of the system, we are using this feature to detect where the coverage of the system must be extended.

Finally, explicit knowledge can be exchanged with external parties like other hospitals. The representation within the system is already the transfer format for information. Semantic Web technologies are by design open for the integration of knowledge that is relative to different ontologies and rules. Therefore we intend to use mainly such technologies for the realization of the retrieval system.

2.2 Use cases and technical requirements

The technical analysis and design of the pathology retrieval system is closely related to typical usage scenarios, which are not necessarily related to routine pathology. Most probable, the system will be used for under-diagnosed cases, where a second or third opinion is to be consulted or the specialist usually reverts to certified control sources, like Internet or printed material. Such information sources have an essential drawback: they offer limited capabilities for a thematically focused search. Both manual search within printed materials and Internet search, based on common or even medicine-related search engines, is time-consuming and not specific enough to be integrated in everyday pathology. Instead, our system offers the possibility to search the archive of medical reports for similar cases or differential diagnosis (retrieve findings with similar symptoms, but different diagnosis). It is improbable that the system will be consulted for routine cases, covering approximately 80 percent of the total archive, which are on the fly analyzed by the pathologists without the need for additional information sources.

The acceptance of the system is strictly related to its "minimal invasive" character: it should not imply any change of the current work flows [1] and should achieve good precision results. Recall is also important, but since the two parameters are usually contradictory, we favor precision, mainly because of the predominant usage of the system for under-diagnosed cases, within which every detail may play an important role for the final results. The minimal invasive feature will be reflected in a careful design of the user interfaces and a intuitive query language.

Another important setting is teaching: therefore, the system should be able to generate different reference materials and to retrieve information about typical pathology cases and their diagnosis. The key feature for the second scenario is the flexibility to generate and present domain information.

The network aspect is important for both settings. Pathologists use the system for cases where they need the remote collaboration of other specialists. The teaching scenario assumes also a distributed infrastructure, so that the resources can be accessed anytime, anywhere. The usage of Semantic Web technologies on one side, and of standards like XML/OWL and the medical HL7/DICOM are conditions for the realization of this requirement.

Scalability and performance are critical factors for the acceptance of retrieval system. In our application, the amount of image data is impressive. Every particular case contains up to 10 medical reports. Each of these are based on up to 50 digital histological images, each of them with a size of 4-5 GB. Our first prototypical implementation of the system will deal with approximately 400 reports and a part of the corresponding digitised slides. The storage of images will still be subject to the use of specialized image databases. Our approach of resorting to the description of images contained in the reports and their processing in the

[1] The system should be integrated to available digital pathology projects, like the Digital Virtual Microscope (see Section 1.1)

system makes the requirements on scalability with the number and complexity of cases independent on the size of the image data. There is no image processing foreseen, instead we use the result of the image analysis performed by human experts, the pathologists. Remaining scalability and performance issues are affected by the quality of the underlying Semantic Web components and the complexity of models used and inferences drawn therein. Currently, there are strong effort to produce Semantic Web components with industrial strength, such as inference engines that go beyond the poor performance of early research prototypes. Our system will benefit from this performance gain in the infrastructure. The complexity of models, rules and queries triggering inferences remains a critical issue. While we have a substantial basis of models with existing standards it it not clear yet, what heuristics should guide the selection of the granularity of models eventually used and of the details of rules applied when report "similar" cases. Therefore we restrict ourselves to small models and rulesets that generate a suffient precise answers by the system with minimal inferencing effort. The precise methodology for doing so is presented in 3.2.

3 Engineering the System

Technically the system resorts to Semantic Web technologies. The Semantic Web ([1]) aims to provide automated information access based on machine-processable semantics of data. The final vision is to develop a technological framework that will transform the Web in an huge network of both human- and machine-understandable knowledge with various specialized reasoning services. The first steps in this direction have been made through the realization of appropriate representation languages for Web knowledge sources like RDF(S) and OWL and the increasing dissemination of ontologies, that provide a common basis for annotation and support automatic inferencing for knowledge generation.

Our approach makes use of these Semantic Web technologies in order to represent pathology knowledge explicitly and, consequently refine the retrieval algorithms on a semantic level: medical and generic ontologies are integrated into a pathology knowledge base, which serve also as annotation vocabulary for medical reports and histological images. We use OWL/RDF(S) for the representation of the knowledge base and for the annotation of the information items and XML-based medical standards like HL7/CDA ([10, 9]) for the reports.

In medicine and biology exhaustive domain ontologies have been developed and are constantly incorporating new pieces of knowledge. Ontologies like UMLS ([18]), GALEN ([6]), Gene Ontology ([4]) provide a good basis for the development of Semantic Web applications for medicine purposes. These ontologies are therefore used as the initial knowledge base of the semantical retrieval system for pathology. In addition, to put our goals into practice we still need to integrate the individual domain knowledge sources and to adapt them to the requirements of the Semantic Web, which means in the first place to formalize them in a Semantic Web representation language. We address the main issues w.r.t. available medical ontologies and the concrete formalization in detail in Section 3.2

3.1 System architecture

We propose the following system architecture, which has arisen from the use cases and the corresponding technical requirements (Figure 1):

- description component
- knowledge component
- transformation component
- application components

In the following we briefly explain the role of each component and their interaction. The core of the system architecture is the knowledge component, which consists of domain and generic ontologies, as well as a rule engine. The description component allows the XML encoding of the textual and pictural data. Both the available pathology data base at the Charité hospital and data to be generated are described in XML in this manner. The transformation component takes the XML-structured data set, analyzes it linguistically and semantically and integrates it within the semantic network underlying the knowledge component. Due to the application-oriented caracter of the system, special attention in the architecture is paid to the application components, which implement the functionality of the system as presented in Section 2. The search component is used both by pathologists in order to retrieve information concerning diagnosis tasks or by teaching personnel and students. We plan also a component for the generation of statistical evaluations (e.g. related to the most frequent disease symptoms, relationships between patient data and disease evolution etc.) and for the generation of case-oriented teaching materials and presentations (see Figure 1). The quality checking service is intended to evaluate the consistency of medical reports.

3.2 Main components

The Description Component The description component is concerned with the basic formalization of medical reports and digital histological images. For this purpose it deals with two principal data sources: data, which is already available at the Institute of Pathology at the Charité hospital and future data. The goal of this process is to offer a homogeneous encoding of medical reports, on one side and picture annotations on the other side, both for existent and future material. The data should be first encoded in XML and subsequently analyzed using ontology-enhanced text analysis algorithms in order to be annotated with ontology concepts. For the generation of new XML-based information we developed an editor tool, which can be integrated in the actual version of the Digital Virtual Microscope ([20, 15]). By means of this tool pathologists can analyze digitized histological images and simultaneously enter or update the corresponding medical report, which are subsequently stored in a XML data base. The second source of raw data was naturally the medical reports archive at the Charité. The medical reports of this type have been extracted from their primary text-oriented storage and transformed in XML.

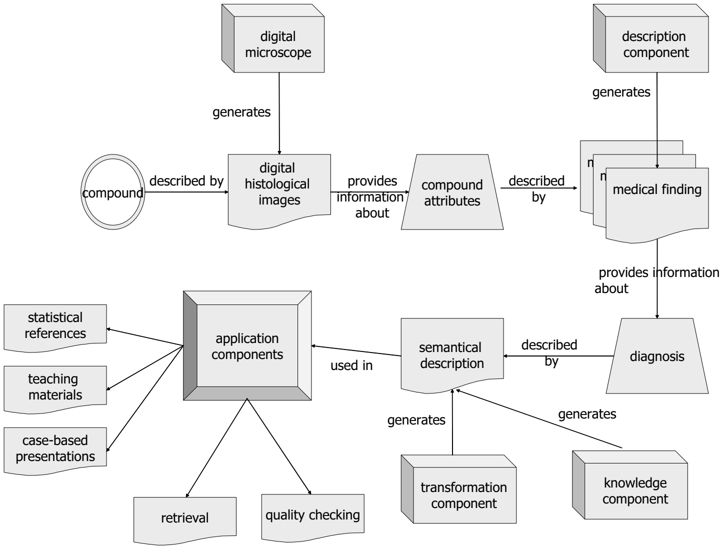

Fig. 1. System architecture "Semantic Web for Pathology"

We developed a HL7/CDA compatible XML-scheme for the medical reports, which reflect the logical structure of the data. Such medical data is organized more or less consequently in four major parts:

- **macroscopy** describing physical properties and the appearance of the original compound.
- **microscopy** concerned with the detailed description of the slides analyzed at the microscope.
- **diagnosis** summarizing the conclusions and the diagnosis
- **comments** usually presenting additional facts playing a role in the diagnosis argumentation (patient data, patient history etc.) or an alternative diagnosis for ambiguous cases.

Besides, such a medical report contains also information from the patient record and references to digital images. The connection to the digital images is fundamental for an efficient retrieval, which should contain apart from the relevant textual information the corresponding image region the pathologist refers to in a certain portion of text. Since the size of such images is 4-5 GB, it is not sufficient to retrieve complete images to a certain user query, but the concrete image sector. For this purpose we use the functionality of the virtual digital microscope, which allows digital slides to be annotated with so-called "observation paths" on one side, and registry an additional "dictation path". The observation path contains image coordinates, image resolution and time stamps registered while the pathologist was analyzing a specific digital image. The dictation path sums up the same data, this time registered while the pathologist was typing the medical report. The complete path-related information flows in the “diagnosis path”, which mirrors the way the diagnosis decision was accomplished. A fragment of a XML-encoded medical report is presented in Figure 2

The proposed XML-Scheme reconstructs the structure of the real medical reports and is HL7-compatible. Though the compatibility restricts the format of the XML reports (the information must be encoded within “section”, “paragraphs” and “coded_entry” tags, which is not necessarily the most straight forward manner of formalizing it), it is an important issue, especially for the distributed setting, for the exchange and reuse of information.

The Knowledge Component The knowledge component includes the medical knowledge base and the algorithms for the realization of the applications. As mentioned in Section 3.1 it is build of a library of domain and generic ontologies, a rule engine and the annotated pathology data.

As input for the medical knowledge base we use UMLS ([18]), as the more complex medical thesaurus currently available. UMLS as in the actual release contains over 1,5 million concepts from over 100 medical libraries and is permanently growing. New sources and actual versions of already integrated sources are permanently mapped to the UMLS knowledge format. Due to the complexity of the thesaurus and the limitations of current Semantic Web tools we need to customize the available medical collection w.r.t. to two important axes: the identification of relevant libraries and concepts corresponding to “lung pathology” from UMLS and their adaption to the particularities of language and vocabular of the case report achive.

This two-phase approach is justified by the application-oriented character of the system. We do not intend to build a general Semantic Web knowledge base for pathology, or even lung pathology, but one, which is tailored for and can be efficiently used in our application setting. Despite standards and tools for the main technologies, building concrete Semantic Web applications, their potential and acceptance at a larger scale is still a challenging issue for the Semantic Web research community.

Identifying relevant knowledge in UMLS The straight-forward method to address this issue is to use the UMLS Knowledge Server ([19]), which provide the MetamorphoSys tool ([18]) and an additional API to tailor the thesaurus to specific application needs. However, both allow mainly syntactical filtering methods (e.g. exclude complete UMLS sources, exclude languages or term synonyms) and do not offer means to analyze the semantics of particular libraries or to use only relevant parts of them. We adopted two approaches to overcome this problem.

- *Top-down Approach* The aim of the top-down approach was to restrict the huge amount of medical information from UMLS to the domain “pathology”. For this purpose, we consulted a team of domain experts (pathologists), who identified potential relevant UMLS libraries. However, the complexity and content heterogeneity (most of the libraries contain concepts belonging

```
<?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>
...
<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="SKFX" /></swpatho:swpatho-ssu>
<swpatho:swpatho-ssu type="E-Nummer" country="de" version="v1">
<swpatho:E-Nummer V="E01152-01" /></swpatho:swpatho-ssu>
</local_header>
...
</clinical_document_header>
<body>
  <section><caption>Befund</caption>
    <section><caption>Makroskopie</caption>
       <paragraph><content>Zwei Gewebszylinder von 15 und 4 mm Laenge.
       </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>Stanzbiopsate aus Lungengewebe mit
      deutlicher Stoerung der alveolaren  Textur, soweit noch
      nachweisbar deutlich Verbreiterung der Alveolarsepten,
      stellenweise Nachweis von Bronchialepithelregeneraten.
      Restliche Alveolarlumina z.T. durch Fibroblastenproliferate
      verlegt. Im Interstitium ein gemischt entzuendliches Infiltrat,
      bestehend aus Plasmazellen und Lymphozyten. Darunter
      relativ viele CD3-positive kleine und mittelgrosse
      T-Lymphozyten und CD68-positive Makrophagen.</content></paragraph>
   </section>
   <section><caption>Kritischer_Bericht</caption>
       <paragraph><content>Stanzbiopsate aus der Lunge mit Zeichen der
       organisierenden Pneumonie (klin. Mittellappen).</content></paragraph>
   </section>
   <section><caption>Kommentar</caption>
      <paragraph><content>Nach klinischer Angabe vordiagnostiziertes
       kutanes T-Zell-Lymphom, jetzt 2 bis 3 cm grosse
       pleurastuendige Raumforderung im Mittellappen.
       Im vorliegenden Material kein Anhalt fuer eine
       Lymphom-Manifestation. Kein Karzinom.</content></paragraph>
   </section>
  </section>
</body>
</levelone>
```

Fig. 2. Fragment of a medical report in XML

to different medicine specialities) of the particular libraries made a manual identification difficult and inefficient. Approximately 50 percent of the UMLS libraries have been selected as possible relevant for lung pathology, containing more than 500000 concepts. Managing an ontology of such dimensions with Semantic Web technologies is related to unsolved issues w.r.t. to scalability and performance of the system. Besides, building the knowledge base implies also a subsequent adaptation of the content, performed by domain experts, that should be able to evaluate and modify the ontology. Therefore, besides technical drawbacks, an ontology of such complexity can not be used efficiently by humans as well.

- *Bottom-up Approach* In the second approach we used the case reports archive to identify concepts, which actually occur in medical reports (i.e. are really used by pathologists while putting down their observations and therefore will also occur as search parameters). For this purpose we used a retrieval engine mapping a lexicon representing the vocabulary of the reports archive to the content of the UMLS sources. The lexicon containing the most frequent nominal phrases was the result of the lexical analysis of the medical reports (in German). The lexicon was subsequently translated to English (due to the restricted set of German terms within UMLS: e.g. from 500000 concepts only 12000 have corresponding German translations in the actual version 2003AC of UMLS) and compared to UMLS. The result of this task was a list of 10 UMLS libraries, still containing approximately 350000 different concepts. The size of the concept set can be explained if we consider the fact that the UMLS knowledge is concentrated in several major libraries (e.g. MeSH([11]), SNOMED98 ([17])), which cover important parts of the complete thesaurus and therefore contain the most of the concepts in our lexicon. To differentiate among the derived libraries we mapped in a second step 10 central concepts in lung anatomy and extract similar or related concepts from UMLS Sources. A total of approximately 400 concepts describing the anatomy of the lung served as initial input for the domain ontology.

Adapting the ontology to the application domain The linguistic analysis of the patient report corpus evidenced the content-related limitations of UMLS. Comparing the results of the lexical analysis with the UMLS content, we recognize some possible extension directions for our ontology: properties like solid, colour, unary predicates and generic properties and concepts: length, diameter, space, spacial objects and their relations. Domain-specific extensions are also revealed through a comparison of the corpus-based lexicon and the generated ontology. For this purpose we modelled additional pathology-specific concepts, like the components and typical core content of a medical report, and integrate them in the available ontology.

OWL Representation The next important issue after identifying an initial set of relevant concepts is the transformation of the raw UMLS data in a Semantic Web formalization, like RDF(S) or OWL. Our analysis in the application domain has revealed the necessity of a powerful representation language, which can capture most of the semantical features of the medical knowledge. For this purpose we will use mostly OWL instead of RDF(S) because of its expressiveness and inferencing capabilities ([6]).

Medicine ontologies though containing a huge amount of concepts or termini have seldom been developed for machine processing, but rather as controlled vocabularies and taxonomies for specific tasks in medicine ([14]). UMLS distinguishes between two data models, which are closely interrelated: The UMLS Semantic Network, containing generic medicine concepts ("semantic types") and relations ("semantic relations"), and the UMLS Metathesaurus. The last one incorporates libraries, like Gene Ontology, SNOMED or MeSH, and consists of "UMLS concepts" referencing semantic types and UMLS relations partially mapped to semantic relations. From a strict Semantic Web point of view UMLS proved to be deficiently designed and incomplete. Apart from the absence of an at least Semantic Web compatible representation language, it adopt an error-prone modeling style, which is characterized by few semantic relations among concepts and an ambiguous way to interpret such relations (e.g. concepts of the UMLS Metathesaurus are connected through relations like "related", "broader", "narrower", "similar", "other-related"). A typical example is the usage of the relation "is-a" for both instantiation and specialization/generalization, the usage of a unique "part-of" relation with different meanings ("functional part", "content", "component", "substance") or the usage of one of these relations instead of the other. Mathematical properties of the same semantical relation (e.g. transitivity) are not fulfilled for each pair of concepts connected by the relation and the "is-a" relation between two concepts does not always guarantee the inheritance of the properties of the parent concept to its children. Another "challenging" modeling issue are so-called "blocked" relations in the UMLS Semantic Network, which have the property that they are not inherited along "is-a" hierarchies. Besides relations, UMLS contains a huge set of conceptual entities, organized in several taxonomies. The classification criteria for concepts are inconsistent and incomplete. Different, unspecified granularities are used within a hierarchy and properties may not be inherited along inheritance paths.

We generated a core domain ontology in OWL based on the original UMLS knowledge base. From a modelling perspective, we modell each UMLS concept as an OWL class, save associated definitions and alternative concept names (so-called UMLS Terms and UMLS Strings) with language specification (German and English) and related it to the corresponding UMLS sources. We also map UMLS relations with a specified meaning to range restrictions on the corresponding concepts and cumulate fuzzy relations like "synonyms", "related", "other-related" etc. to a generic "related_to" relationship. We leave additional details about the modelling primitives to another paper. This way the generated ontology is the result of a partial direct mapping from the UMLS thesaurus. The UMLS Semantic Network was also formalized in OWL since every UMLS concept is connected to it. After an automatic discovery of the (logical) inconsistencies of the modell, we are currently work at a methodology for the semi-automatic adaptation of the OWL ontology in order to correct these errors and to include pathology-specific knowledge, like the structure of case reports and frequently-used concepts from texts not supported by UMLS.

The Transformation Component The transformation component is closely related to the knowledge component and implements features required for the text-based processing of the medical reports and image descriptions. For this purpose we are developing a noun phrasing module, which identifies domain-specific phrases from medical reports. The modules incorporates a tokenizer, a tagger and a ontology-based phrase generator. The phrase generation process interacts with the knowledge base, since it uses medical ontologies to identify relevant (multi-word) phrases and in the same time puts together a lexicon, tailored for the particular application setting: like the domain of lung pathology and the language used in the medical reports, which is German. The lexicon provides also indications about the usage limitations of an essentially English-oriented thesaurus like UMLS in our concrete setting. The case reports are annotated by means of text processing with concepts from the knowledge base/ontology. Therefore the linguistic component needs to recognize concepts and their characteristics, relations among concepts from text. The result of these procedure is an intermediate logical representation (see Figure 3 for an example representation of the XML fragment report from Figure 4). As a result of the phrasing module, the XML-encoded medical reports contain semantic relevant phrases, which can be referenced to concepts of the knowledge base.

The logical forms produced by the parser are transformed into OWL-compliant representations. This process is fairly straightforward, as should be clear from comparing the intermediate representation in Figure 4 with the target representation in Figure 5.[2]

- unique identifiers for the instances of concepts have to be created, and
- in cases of plural entities ("two cylinder" → $card(x, 2)AND_{:}cylinder(x)$), several separate instances have to be created.
- Appropriateness conditions for properties are applied: if a property is not defined for a certain type of entity, the analysis is rejected.

Note that this also handles potential syntactic ambiguity, since it might filter out analyses on the grounds because they specify inconsistent information.

Application Components The Semantic Web for Pathology will assist the following application components:

- **search component** will be used primarily for diagnosis tasks. It will allow not only the basic retrieval of text/image information items, but also support differential diagnosis tasks. The semantic retrieval is oriented towards several typical categories of queries:

[2] Every medical report will be formalized in OWL as instances. The corresponding concepts are modelled separately (as classes).

```
<section><caption>Befund</caption>
    <section><caption>Makroskopie</caption>
       <paragraph><content>[1]Zwei Gewebszylinder von 15 und 4 mm Laenge[1].
       </content></paragraph>
    </section>
    <section><caption>Mikroskopie</caption>
     ...
      <paragraph><content>[2]Stanzbiopsate aus Lungengewebe mit
      deutlicher Stoerung der alveolaren  Textur, soweit noch
      nachweisbar deutlich Verbreiterung der Alveolarsepten,
      stellenweise Nachweis von Bronchialepithelregeneraten[2].
      [3]Restliche Alveolarlumina z.T. durch Fibroblastenproliferate
      verlegt[3]. [4]Im Interstitium ein gemischt entzuendliches Infiltrat,
      bestehend aus Plasmazellen und Lymphozyten[4]. [5]Darunter
      relativ viele CD3-positive kleine und mittelgrosse
      T-Lymphozyten und CD68-positive Makrophagen[5].</content></paragraph>
   </section>
   <section><caption>Kritischer_Bericht</caption>
       <paragraph><content>[6]Stanzbiopsate aus der Lunge mit Zeichen der
       organisierenden Pneumonie (klin. Mittellappen)[6].</content></paragraph>
   </section>
   <section><caption>Kommentar</caption>... </section>
  </section>
```

Fig. 3. Input of the transformation component

```
[1]card(x1, 2) AND cylinder(x1) AND length(x1, [15, 14])
[2]unspec_plur_det(x2) AND punch_biopsat(x2)
       AND from_rel(x2, x3) AND unspec_plur_det(x3) AND lung_tissue(x3)
       AND with_rel(x3, x4) AND def_det(x4) AND disturbance(x4, x5)
       AND def_det(x5) AND texture(x5) AND alveolar(x5)
   unspec_det(x6) AND extension(x6, x7) AND def_det_plur(x7)
       AND aleveolar_septum(x7) AND unspec_det(x8) AND evidence(x8, x9)
       AND indef_det(x9) AND epithelial(x9) AND bronchial(x9) AND regenerates(x9)
[3]def_det(x10) AND alveolarlumina(x10)
   unspec_det_plur(x11) AND fibrolastial_proliferate(x11)
[4]def_det(x12) AND interstitium(x12)
   indef_det(x13) AND inflammatory(x13) AND infiltrate(x13) consisting_of_rel(x13, x14)
       AND unspec_det_plur(x14) AND konj(x14, x15, x16) AND plasma_cell(x15)
       AND lymphocyte(x16)
[5]indef_det_plur(x17)  AND konj(x17, x18, x19) AND  t_lymphocyte(x18)
       AND cd68_positive(x19) AND macrophagus(x19)
[6]indef_det_plur(x20) AND punch_biopsate(x20) AND from_rel(x20, x21)
       AND def_det(x21) AND lung(x21) AND with_rel(x20, x22) AND evidence(x22, x23)
       AND def_det(x23) AND organising(x23) AND pneumonia(x23)
```

Fig. 4. Intermediate output of the transformation component

```
....
 <Lung_Tissue rdf:ID="lung_tissue_x3">
    <partOf>
      <Lung_C0024109 rdf:ID="lung1">
        <hasSource rdf:resource="umlssources.owl#UWDA"/>
        ... properties of the lung ...
      </Lung_C0024109>
    </partOf>
  </Lung_Tissue>
   <Punch_biopsat rdf:ID="punch_biopsat_x2">
    <from rdf:resource="#lung_tissue_x3"/>
  </Punch_biopsat>
  <alveola rdf:ID="alveola_x5">
    <hasTexture rdf:datatype="http://www.w3.org/2001/XMLSchema#string">disturbed</hasTexture>
    <relatedTo rdf:resource="#lung1"/>
  </alveola>
  <Cylinder rdf:ID="cylinder_x1">
    <length rdf:datatype="http://www.w3.org/2001/XMLSchema#float">    15.0</length>
    <formOf rdf:resource="#punch_biopsat_x2">
  </Cylinder>
  <Cylinder rdf:ID="cylinder_x2">
    <length rdf:datatype="http://www.w3.org/2001/XMLSchema#float">    14.0</length>
    <formOf rdf:resource="#punch_biopsat_x2">
  </Cylinder>
....
```

Fig. 5. Fragment of the OWL output of the transformation component

- **statistical queries** e.g. the probability/frequency of a particular carcinoma in a certain age group.
- **matching queries** e.g. comparison of cases with common characteristics, text and image information to similar cases.
- **image queries** e.g. cases containing images with certain content- or image-specific constraints.

Besides, the retrieval should be adapted to the characteristics of the pathology domain and involve issues like the diagnosis path. (see Section 3.2).

- **quality checking component** will be used in quality assurance and management of diagnosis processes. Quality criteria, diagnosis standards and their verification are expressed by means of rules.
- **statistical component** will generate statistical material related to the relative frequency or demographic distribution of diseases and their complications.
- **teaching component** will generate teaching materials, using features of the previous components (statistical studies, reference cases)

4 Related Work

Medicine is one of the best examples of application domains where ontologies have already been deployed at large scale and have already demonstrated their utility. Most of these domain ontologies (UMLS inclusively)underlie different design requirements as computerized or more specific Semantic Web applications. They are actually huge collections of medical terms, organized in hierarchies and cannot be used directly in Semantic Web applications. This issue has been addressed in project GALEN ([6]), where the authors developed a special representation language, tailored for the particularities of the (English) medical vocabulary. However, the usage of a proprietary representation makes the ontological knowledge difficult to be extended by third parties or exchanged in a Semantic Web.

The usage of ontologies for building knowledge bases for medicine has already been subject of several research projects ([2, 13, 7, 3, 8]). The most important representatives are the ONIONS ([7]) and MEDSYNDIKATE ([13]) projects. In ONIONS the authors aim to develop a generic framework for ontology merging and use UMLS as an example to apply their methodology. Therefore they need a detailed analysis of the ontological properties of UMLS, using a Loom formalization. MEDSYNDIKATE is also confronted with the ontological commitment beyond UMLS in order to use it in text processing algorithms for knowledge discovery. UMLS serves in this case as an annotation vocabulary for medical texts. Both projects offer valuable experiences and facts concerning UMLS and medical ontologies generally, but they do not use Semantic Web technologies to facilitate knowledge share and reuse, which is the crucial feature of ontologies. An interesting approach can also be found in [2], where the authors compare UMLS with other ontologies (e.g. WordNet ([12], GeneOntology) to establish its appropriateness as terminology for biomedical applications.

5 Conclusions and Future Work

In this paper we presented our work towards a Semantic Web based retrieval system for pathology. The system is based on a comprehensive knowledge base, which formalizes pathology-relevant knowledge explicitly by integrating available medicine ontologies like UMLS and rules describing diagnostic guidelines. It is intended to provide both retrieval and knowledge management functionalities. In order to achieve these goals we designed by now the system architecture, adopted XML-based schemes for the uniform representation of medical reports and digital images and developed a methodology for the construction of the pathology knowledge base. We generated a prototype ontology for lung pathology based on UMLS knowledge sources and the lexical analysis of an archive of pathology medical reports. Current work includes the specification of algorithms for the semantical annotation of the medical reports, for the incremental enrichment of the core ontology and for the acquisition of domain-specific rules.

Acknowledgment The project “Semantic Web in the Pathology” is funded by the Deutsche Forschungsgemeinschaft, as a cooperation among the Charité Institute of Pathology, the Institute for Computer Science at the FU Berlin and the Department of Linguistics at the University of Potsdam, Germany.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila. "The Semantic Web". *Scientific American*, 284(5):34–43, 5 2001.
2. A. Burgun and O.Bodenreider. Mapping the UMLS Semantic Network into General Ontologies. In *Proc. of the AMIA Symposium*, 2001.
3. G. Carenini and J. Moore. "Using the UMLS Semantic Network as a Basis for Constructing a Terminological Knowledge Base: A Preliminary Report". In *Proceedings of 17th Symposium on Computer Applications in Medical Care (SCAMC '93)*, 1993.
4. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–30, 2000.
5. F. Demichellis, V. Della Mea, S. Forti, P. Dalla Palma, and C.A. Beltrami. "Digital storage of glass slide for quality assurance in histopathology and cytopathology". *Telemed Telecare*, 8(3):138–42, 2002.
6. Ontology GALEN. http://www.opengalen.org, 2001.
7. 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.
8. H. Gu, Y. Perl, J. Geller, M. Halper, L. Liu, and J. Cimino. "Representing the UMLS as an OODB: Modeling issues and advantages", 2000.
9. HL7 Standard. http://puck.informatik.med.uni-giessen.de/people/messaritakis/-hl7xml/hl7stand.htm, 2000.
10. The HL7/CDA Standard. http://www.hl7.org, 2000.
11. Medical Subject Headings. http://www.nlm.nih.gov/mesh/meshhome.html, 2003.
12. G. A. Miller. "WordNet: a lexical database for English". *Communications of the ACM*, 38(11):39 – 41, 1995.
13. 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.
14. S. Schulz, M. Romacker, and U. Hahn. "Knowledge engineering the UMLS". *Stud Health Technol Inform*, 77:701–5, 2000.
15. Patentanmeldung: Slide Scanner – Vorrichtung und Verfahren, 2002. Aktenzeichen 102 36 417.6 des DPMA vom 5.8.2002.
16. J. Slodkowska, K. Kayser, and P Hasleton. "Teleconsultation in the Chest Disorders". *Eur J Med Res*, 7 (Suppl I):80, 2002.
17. Snomed International. http://www.snomed.org, 1998.
18. Unified Medical Language System. http://www.nlm.nih.gov/research/umls, 2002.
19. UMLS Knowledge Source Server. http://umlsks.nlm.nih.gov, 2003.
20. Patentanmeldung: Virtuelles Mikroskop – Vorrichtung und Verfahren, 2002. Aktenzeichen 102 25 174.6 des DPMA vom 31.05.2002.