Pillars of Ontology Treatment in the Medical Domain

Daniel Sonntag, DFKI German Research Center for AI
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
phone: +49 681 3025254; fax: +49 681 3025020; e-mail: sonntag@dfki.de

Pinar Wennerberg (Siemens),

Paul Buitelaar, DFKI German Research Center for AI

Sonja Zillner (Siemens).

Pillars of Ontology Treatment in the Medical Domain

In this chapter we describe the three pillars of ontology treatment in the medical domain in a comprehensive case study within the large-scale THESEUS MEDICO project. MEDICO addresses the need for advanced semantic technologies in medical image and patient data search. The objective is to enable a seamless integration of medical images and different user applications by providing direct access to image semantics. Semantic image retrieval should provide the basis for the help in clinical decision support and computer aided diagnosis. During the course of lymphoma diagnosis and continual treatment, image data is produced several times using different image modalities. After semantic annotation, the images need to be integrated with medical (textual) data repositories and ontologies. We build upon the three pillars of knowledge engineering, ontology mediation and alignment, and ontology population and learning to achieve the objectives of the MEDICO project.

INTRODUCTION

Clinical care and research increasingly rely on digitized patient information. There is a growing need to store and organize all patient data, such as health records, laboratory reports and medical images, so that they can be retrieved effectively. At the same time it is crucial that clinicians have access to a coherent view of these data within their particular diagnosis or treatment context.

With traditional applications, users may browse or explore visualized patient data, but little to no help is given when it comes to the interpretation of what is being displayed. This is due to the fact that the semantics of the data is not explicitly stated, which therefore remains inaccessible to the system and therefore also to the user. This can be overcome by the incorporation of external medical knowledge from ontologies which provide the meaning (i.e., the formal semantics) of the data at hand.

Our research activities are in the context of the THESEUS MEDICO project. MEDICO addresses the need for advanced semantic technologies in medical image and patient data search. The objective is to enable a seamless integration of medical images and different user applications by providing a direct access to image semantics. A wide range of different imaging technologies in various modalities exist, such as 4D 64-slice Computer Tomography (CT), whole-body Magnet Resonance Imaging (MRI), 4D Ultrasound, and the fusion of Positron Emission Tomography and CT (PET/CT). All these image modalities have the common property that their semantic contents include knowledge about human anatomy, radiology, or diseases.

One important requirement for advanced applications in semantic image retrieval, clinical decision support and computer aided diagnosis is the comparative exploration of similar patient information. For this purpose, we envision a flexible and generic image understanding software for which semantics of the images plays the major role for access and retrieval. However, currently, large amounts of medical image data are indexed by simple keywords to be stored in distributed databases without capturing any semantics.

The objective of MEDICO is to build the next generation of intelligent, scalable and robust search engines for the medical imaging domain, based on semantic technologies. With the incorporation of higher level knowledge represented in ontologies, different semantic views of the same medical images (such as structural aspects, functional aspects, and disease aspects) can be explicitly stated and integrated. Thus, the combination of formal semantics with image understanding helps building bridges between different but related domains that can be used for comparative exploration of patient data. MEDICO is a consortium research project funded by the German Federal Ministry of Economics with several R&D sites and the Erlangen University Hospital as a clinical partner.
Visit http://theseus-programm.de/scenarios/en/medico.

Within the MEDICO project, one of the selected scenarios aims for improved image search in the context of patients that suffer from lymphoma in the neck area. Lymphoma, which is a type of cancer affecting the lymphocytes, is a systematic disease with manifestations in multiple organs. During the course of lymphoma diagnosis and continual treatment, image data is produced several times using different modalities. As a result, the image data consist of many medical images in different formats, which additionally need to be associated with the

corresponding patient data. Hence, the lymphoma scenario is particularly suitable to demonstrate the strength of a semantic search engine as we envisioned in MEDICO.

To address the challenges of advanced medical image search, different medical resources need to be semantically integrated. Consequently, the following four research questions arise:

1) How is the workflow of the clinician, i.e.,

    a) What kind of information is relevant for his daily tasks?

    b) At what stage of the workflow should selected information items be offered?

2) What are the particular challenges and requirements of knowledge engineering in the medical domain?

    a) Can those challenges be addressed by a semi-automatic knowledge extraction process based on clinical user interactions?

    b) Can we embed the semi-automatic extraction process into the clinician's workflow?

3) How can different possibly overlapping data sources (i.e., ontologies) be aligned?

4) How can we learn and populate ontologies?

MEDICO's vision of the semantic medical search relies on ontology-based annotation of medical images and the related patient data. This allows us to mark-up the content at a higher level of granularity that goes beyond simple keywords. To realize this, the use of metadata from multiple, disparate but nevertheless related ontologies is required.

We will describe the three pillars of ontology treatment in the medical domain in a comprehensive case study within MEDICO. These pillars are knowledge engineering, ontology mediation and alignment, and ontology population and learning. We build upon these pillars to achieve the objectives of MEDICO.

The contribution of this book chapter is this description of the pillars of ontology treatment in the medical domain and the overview of our implementations of these pillars. For example, the approach for realizing a medical image search scenario based on semantic technologies within an industry setting represents one of the pillars (knowledge management). We put the focus on the challenges, requirements, and possible solutions related to ontology alignment.

The remainder of this book chapter is organized as follows. Section 2 outlines the pillars of ontology treatment. Section 3 describes our implementations of the knowledge engineering requirements of a clinician in the context of his daily work along three clinical scenarios. We will also discuss the medical knowledge engineering workflow. Section 4 addresses the challenges and possible solutions for mediating and (semi-automatically) aligning different medical ontologies. In section 5 we discuss and analyze how MEDICO ontologies can be populated in a semi-automatic way. The final section concludes and describes our future work in the THESEUS MEDICO use case.

PILLARS OF ONTOLOGY TREATMENT IN THE MEDICAL DOMAIN

According to the clinical knowledge requirements, we can identify three pillars of ontology treatment in the medical domain. These pillars should allow us to improve the clinical reporting process, the patient follow-up process, and the clinical disease staging and patient management process. This is achieved by the use of metadata from multiple, related medical ontologies. In the following, we will describe the pillars of ontology treatment: knowledge engineering; ontology mediation and alignment; and ontology population and ontology learning.

Knowledge Engineering:

What is the recommended medical information management and ontology engineering process and what semantic-driven recommendations can be given to enhance existing medical knowledge repositories? Which recommendations can support building up new medical knowledge repositories? A knowledge engineering methodology (KEMM) helped us to formalize these requirements. How this relates to the doctor's practical interest in using a semantic search engine or dialogue interface is one major part of the practical case study. For example, consider a radiologist at his daily work: The diagnostic analysis of medical images typically concentrates around three questions: i) what is the anatomy? ii) what is the name of the body part? iii) is it normal or is it abnormal? To satisfy the radiologist's information need, this scattered knowledge has to be gathered and integrated from disparate dynamic information sources.

Ontology Mediation and Alignment:

Information integration is concerned with access to heterogeneous information sources (in MEDICO: text patient data, medical images, relational databases) to be mediated in order to provide an integrated view of the data. In addition, we have specific information needs that must be satisfied by these information sources (which should be expressed by query patterns defined over a set of ontologies). In medical imaging and MEDICO, a single ontology is not enough to support the required complementary knowledge from different perspectives, for example anatomy, radiology, or diseases. Ontology mediation and alignment is therefore a key aspect of the semantic information integration task in the MEDICO use case.

We investigate linguistic-based, corpus-based, and speech-dialogue-based ontology alignment approaches in the main part of this case study. We will also discuss the methods that are required for interactive and incremental ontology mapping in the MEDICO use case, and their applicability.

Ontology Population:

Given the set of identified relevant and aligned ontologies, one important aspect of our approach is the automatic extraction of knowledge instances (entities, facts) from text data. This data is widely available in the medical domain in the form of patient records as well as scientific articles. The important aspect is the semantic integration of these mainly unstructured data instances with those derived from other resources (medical images, relational databases) through ontology population across ontologies. In this connection, we describe an interactive GUI environment for the medical expert.

1. Knowledge Engineering in the Medical Domain

MEDICO covers a particularly sensitive domain, i.e., human health. In this domain, the reuse of medical knowledge, which is already present in readily available standardized, high quality medical ontologies engineered by domain experts, is crucial.

In our context, we use the term "knowledge engineering" in the sense it is discussed by Grüninger and Uschold (1996). It is refer to *"methods for creating an ontological and computational basis for reuse of product knowledge across different applications within technical domains."* Consequently, we understand ontology treatment (i.e., ontology mediation and ontology population) as specific knowledge engineering tasks.

Various challenges exist in medical knowledge engineering. One challenge is that the knowledge engineer is not familiar with the complex and comprehensive medical terminology in the medical ontologies. As a result, the application domain remains opaque to him and he cannot verify the knowledge engineering process. Other challenges are the size of the medical ontologies, which overwhelms a non-medical person, not to mention the technical challenges of the software engineering process, for example runtimes.

The major challenge, however, is the so-called "knowledge acquisition bottleneck." We cannot easily acquire the necessary medical knowledge that ought to be used in the software application as it is hidden in the heads of medical experts. Our experience with the MEDICO project shows that common interview methods are neither efficient nor effective enough to acquire the domain knowledge (due to misunderstandings in the communication).

Therefore, we view medical knowledge engineering as an *interactive process* between the knowledge engineer and the clinician. The first essential step requires the knowledge engineer to gather and pre-processes available medical knowledge from various resources such as domain ontologies and domain corpora, whereupon the domain expert, i.e., the clinician, evaluates the outcome of the process and provides feedback.

Thus, we address the "knowledge acquisition bottleneck" problem by concerning ourselves with the question how a bottom-up ontology engineering approach can be established based on a data-driven knowledge pre-processing step (that is followed by a user interactive evaluation step). Here, our focus is on the development of semantic-driven recommendations to enhance existing medical knowledge repositories according to KEMM (Knowledge Engineering Methodology in the Medical Domain).

1.1. Clinical Knowledge Engineering Requirements

Today, medical images provide important information for identifying the patient's diagnosis and appropriate treatment. As medical imaging technologies progress and more and more medical details become more clearly visible, it happens quite often that clinicians discover some suspicious or unknown alteration of particular body parts in medical images. In such situations, the most valuable and relevant information for clinicians can be gained by comparing the non-routine results to other but nevertheless "similar" images. By comparing a given image to other scans and records of patients with similar visual symptoms, e.g., an enlargement of the lymph node of the neck, clinicians can learn about the meaning of the unknown alteration in the context of the progress of the disease.
In contemporary, daily hospital work, clinicians can only manually search for "similar" images. After considering the relevant categories of similarity, they subsequently set one filter

after the other. For instance, a clinician first sets a filter for the imaging modality (e.g., CT angiography), the second filter for the procedure (e.g., coronary angiography), and so on. Beside the fact that this approach is quite time-consuming, it is neither possible to formulate complex and semantically integrated search queries, nor can valuable knowledge of external knowledge resources be integrated.

This is the situation we face today. Thus, in intensive discussions with clinicians we analyzed how the use of semantic technologies can support the clinician's daily work tasks. In particular, we discussed the medical case of lymphoma from the perspective of medical imaging and revealed three typical clinical scenarios that are of interest for further analysis of clinical knowledge requirements:

1. The clinical reporting process;

2. The patient follow-up treatment (i.e., monitoring the patient's health condition and the development of the disease);

3. The clinical disease staging and patient management.

Each scenario induces a list of relevant tasks with particular clinical questions to be answered. Each answer is again based on particular medical data that is (or is not) available and that is typically stored in distributed knowledge and data repositories. The three clinical scenarios require the acquisition of various types of domain knowledge:

1. The *clinical reporting process* focuses on the general question "What is the disease?" (or, as in the lymphoma case, "which lymphoma?") To answer this question, *semantic annotations* on medical image contents are required. These are typically anatomical parts such as organs, vessels, lymph nodes, etc. Image parsing and pattern recognition algorithms can extract the low-level image feature information. The low-level information is used to produce higher-level semantic annotations to support tasks such as differential diagnosis.

2. Within the *patient follow-up process,* the clinician's concern is whether his former diagnosis hypothesis is confirmed by the outcome of the treatment or not. In other words, a clinician can only know what he is treating until he sees how the patient responds (Starbucks, 1993). The questions relevant for this scenario are, "Is the drug effective?", "Has the lesion shrunk?", and "Do the symptoms persist?" Therefore, the clinician is particularly interested in finding out if his prior diagnosis hypothesis can be verified or refuted.

3. In the *clinical staging* and *patient management process* the general concern is with the next steps in the treatment process. The results of the clinical staging process influence the decisions that concern the later patient management process. (*External medical knowledge* comes into play, in the sense that the disease staging results need to be mapped onto the standard clinical staging and patient management guidelines.)

To satisfy the radiologist's information need, this scattered information has to be gathered, semantically integrated and presented to the user in a coherent way. Finally, external resources such as medical guidelines or medical recommendations need to be integrated as well in order to achieve compatibility with the standard decision making and management procedures.

1.2. KEMM Methodology

From the knowledge engineering requirements, we derived a knowledge engineering methodology that is specific for the medical domain (Wennerberg, 2008). Consequently, KEMM (Figure 1) defines seven tasks. The initial task, called *Query Pattern Derivation,* supports the communication between the knowledge engineer and the clinician during the knowledge elicitation process. All other tasks (explained further down) support the medical ontology engineering process.

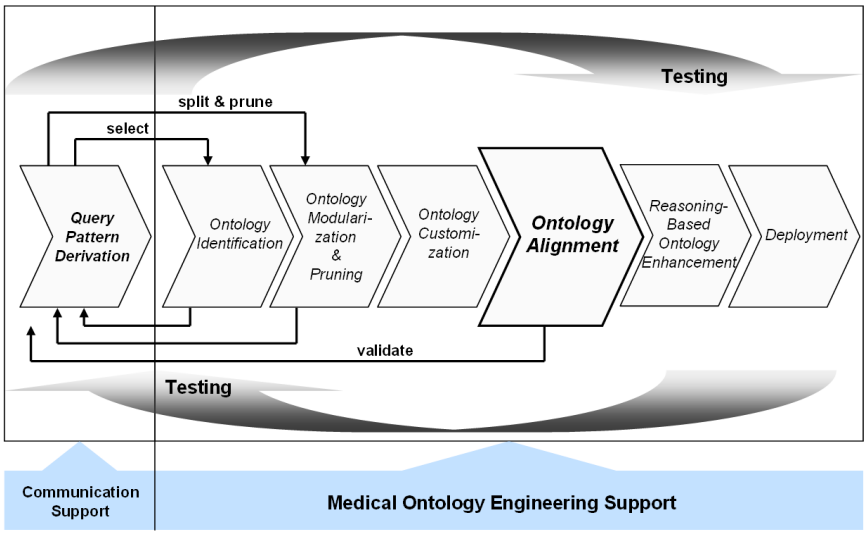


Figure 1: Knowledge Engineering Methodology for the Medical Domain (KEMM) .

Query Pattern Derivation: This task is based on generating a set of hypothetical user queries using domain ontologies and domain corpora that are subsequently evaluated by the clinicians. A combination of various techniques from natural language processing to text mining are employed to derive patterns (described in detail in Buitelaar (2008) and Wennerberg (2008a)). In the MEDICO study, we focused on the patterns for typical clinical queries given the domain ontologies and use case corpora; concept-relation-concept triplets are identified. The pattern derivation can be viewed as a function that takes the domain (sub)ontologies and the corpora as input and returns a partial weighting of the ontologies, whereby the terms/concepts are ranked according to their weights. A complex query pattern example is: (ANATOMICAL_STRUCTURE *located_in* ANATOMICAL_STRUCTURE) **AND** ((RADIOLOGY_IMAGE) Modality *is_about* ANATOMICAL_STRUCTURE) **AND** ((RADIOLOGY_IMAGE) Modality *shows_symptom* DISEASE_SYMPTOM).
The top 4 concepts we identified for the generic query pattern above in the corpora are:

FMA Term	Score/Frequency
lateral	338724
anterior	314721
artery	281961
anterior spinal	219894

RadLex Term	Score/Frequency
x-ray	81901
imaging modality	58682
volume imaging	57855
molecular imaging	57850

Ontology Identification: As the medical image contents essentially relate to human anatomy, radiology, pathology, and/or diseases, we require identifying ontologies from these domains. Consequently, the Foundational Model of Anatomy[i], the Radiology Lexicon[ii], and the NCI Cancer Thesaurus[iii] were set as semantic resources that provide the domain knowledge.

***Ontology Modularization and Pruning**:* Based on these patterns, the ontologies to be reused are identified, pruned, and modularized; the relevant modules are customized and finally integrated. For an effective reuse of the large medical ontologies, we have to construct modular ontology subsets that can be easily navigated by humans and reasoned by machines. The derived set of query patterns determines the criteria for pruning and modularizing the large medical ontologies that were identified in the previous step. These pruned (and modularized) ontologies are then presented to the clinical experts to confirm their relevance and validity.

***Ontology Customization**:* Quite often, the modules extracted from the ontologies have either redundant or missing knowledge; only customized knowledge in terms of domain ontologies meets the requirements with respect to the applications. For example, we defined a relationship *has_nci_code* which relates the concepts in the lymphoma module to the entities in the NCI thesaurus. Another customization was the conversion of the lymphoma related section of the NCI Thesaurus from the flat text format to OWL.

***Ontology Alignment**:* We conceive of ontology alignment as an operation on the extracted ontology modules (rather than the ontologies as a whole). The objective of the alignment is to obtain a coherent picture of separate but related ontology modules. Each customized ontology module represents a piece of knowledge that is necessary to realize the entire application. These knowledge pieces are not arbitrary but they need to be interrelated within the context of the application. The different ontology alignment and mediation approaches will be discussed in more detail in Section 2.

***Reasoning-Based Ontology Enhancement**:* The MEDICO use case is characterized by the reuse and integration of distributed ontological knowledge that may introduce inconsistencies. With the KEMM methodology we concentrate on two specific reasoning services. In our lymphoma use case one objective is to be able to deduce the relevant image modalities (MR, CT scan etc.) given the symptoms of head and neck lymphoma. Via deductive reasoning we target the discovery of valid relationships--spatial, pathological, and physiological--between anatomical structures. ***Testing and Deployment**:* To avoid the propagation of inconsistencies and modeling mistakes, each and every task should be tested for validity, completeness, and coherence.

With the KEMM methodology, our intention is to provide a theoretical framework for the knowledge engineer, whose application domain is healthcare. Based on our experience, we assume that knowledge engineers, who have no or little background knowledge in biomedical sciences, will face similar challenges. Therefore, the goal of KEMM is to inform the knowledge engineers about both domain specific technical challenges and potential communication difficulties with the domain experts. Ontology alignment is the most important ontology treatment pillar in practical terms, and we discuss it next to explain *why* it is necessary. Additionally, we give details on *how* it can be implemented.

2. Ontology Mediation and Alignment

We regard ontology alignment as an important building block of knowledge engineering in the medical domain. Medical knowledge engineering typically requires semantic integration of different medical knowledge, which can be supported by ontology alignment. KEMM demonstrates our view on *when* this should happen within the entire medical knowledge

engineering process. Ontology alignment is an increasingly active research field in the biomedical domain, especially in association with the Open Biomedical Ontologies (OBO)[iv] framework. The OBO consortium establishes a set of principles to which the biomedical ontologies shall conform to for purposes of interoperability. The OBO conformant ontologies, such as the FMA, are available at the National Center for Biomedical Ontology (NCBO) BioPortal[v].

Information integration is concerned with the access to heterogeneous information sources (in MEDICO: text patient data, medical images, and relational databases) to be mediated in order to provide an integrated view of the data. We also have specific information needs to be answered by these information sources, which may be expressed by a query pattern defined over a set of ontologies. As already mentioned, in MEDICO, a single ontology is not enough to provide the required complementary types of knowledge, i.e., the anatomy, radiology, or diseases.

There is a need for clinicians, in particular for radiologists, to be able to have access to coherent information from a single access point. At the center of their search are the medical images of patients, i.e., starting from a specific medical image, the radiologists wish to find all the information that is related to the case. Currently, this is not possible and the radiologist needs to use several systems at different locations.

Semantic annotations can help integrate the related data that is stored in distributed repositories by using commonly agreed annotation vocabularies. Consequently, radiologists can use the same vocabularies (i.e., those used for annotations) for their search and obtain the information from a single access point.

Hence, one of our goals within the context of the MEDICO use case is to offer clinicians and radiologists an integrated view of different kinds of information that is all centered around the medical images. We conceive of a radiology expert as an end user who looks, starting from a certain medical image, for all related information such as patient data, lab reports, treatment plans etc. Obtaining this kind of heterogeneous information from a single access point requires the data to have been previously integrated appropriately. The integration can be achieved while annotating the data with the relevant vocabularies. Nevertheless, during the search the radiologist prefer to use “his vocabulary” (i.e., a radiology specific vocabulary) for convenience. To be able to cover all relevant information by using only one vocabulary as a starting point therefore requires an alignment with other vocabularies that are relevant for image contents and patient data.

De Bruijn et al. (2006) offer terminological clarification for all the related research activities around ontology alignment. Accordingly, the reconciliation of differences between ontologies is defined as *ontology mediation*, whereby *ontology mapping* and *ontology merging* are considered as two specific cases of ontology mediation. In the case of ontology mapping the set of correspondences between different ontologies is not a part of the ontologies themselves. Ontology alignments, in this respect, are the results of the (semi-)automatic discovery of these correspondences in a suitable descriptive format. Others have a slightly different but non-contradictory definition. The difference between ontology mapping and ontology alignment according to Johnson et al. (2006) is that the former deals with the identification of equivalent concepts in multiple ontologies, whereas the latter specifically focuses on making the overlapping concepts in multiple ontologies compatible.

Our goal is to identify and post-process the correspondences between the concepts of different medical ontologies that are relevant to the contents of the medical images. This is how we define ontology mediation and alignment. The following scenario illustrates how the alignment of medical ontologies facilitates the integration of medical knowledge from multiple ontologies which are relevant for medical image contents. Suppose we want to help a radiologist who searches for related information about the manifestations of a certain type of lymphoma on a certain organ (e.g., the liver) on medical images. The three types of knowledge that help him would be about the human anatomy (liver), the organ's location in the body (e.g., upper limb, lower limb, neighboring organs etc.), and whether what he sees is normal or abnormal (pathological observations, symptoms, and findings about lymphoma).

Once we know what the radiologist is looking for, we can support him in his search in that we present him with an integrated view of only the liver lymphoma relevant portions of the patient health records, scientific publications abstracts (such as those of PubMed[vi]) as a reference resource, drug databases, experience reports from other colleagues, treatment plans, notes of other radiologists, or even discussions from clinical web forums. From the NCI Thesaurus we can obtain the information that *liver lymphoma* is the synonym for *hepatic lymphoma*, for which holds:

'Hepatic lymphoma' (NCI term),
'disease_has_primary_anatomic_site' (NCI relation),
'Liver' (NCI term and FMA term),
'Hematopoietic and lymphatic system' (NCI term),
'Gastrointestinal system' (NCI term).

With this information, we can now move on to the FMA ontology to find out that *hepatic artery* is a part of the liver (such that any finding that indicates lymphoma at the *hepatic artery* would also imply the lymphoma at the *liver*). RadLex, on the other hand, informs us that *liver surgery* is a *treatment procedure.* Various types of this *treatment procedure* are *hepatectomy*, *hepatic lobectomy*, *hepatic segmentectomy*, *hepatic subsegmentectomy*, *hepatic trisegmentectomy,* or *hepatic wedge excision,* all of which can be applied to treat the disease.

Consequently, the radiologist, who searches for information about liver lymphoma, is presented with a set of patient health records, PubMed abstracts, radiology images etc. that are annotated using the terminology above. In this way, the radiologist's search space is reduced to a significantly small portion of the information available in multiple data stores. Moreover, he receives coherent data, i.e., images and patient text data that are related to each other, from a single access point without having to log in to several different data stores at different locations. In what follows, we will discuss related work in medical ontology mediation and alignments and we will propose our three approaches for the medial domain, i.e., linguistic-based, corpus-based, and dialogue-based to overcome some of the difficulties.

Johnson et al. (2006) take an information retrieval approach to discover relationships between the Gene Ontology (GO) and three other OBO ontologies (ChEBI[vii], Cell Type[viii], and BRENDA Tissue[ix]). GO ontology concepts are treated as documents, they are indexed using Lucene[x] and are matched against the search queries, which are the concepts from the other three ontologies. Whenever a match is found, it is taken as evidence of a correspondence. This approach is efficient and easy to implement and can therefore be successful with large medical ontologies. However, it does not account for the complex linguistic structure typically observed in the concept labels of the medical ontologies and may result in inaccurate matches.

The main focus of the work by Zhang et al. (2004) is to compare two different alignment approaches that are applied to two different ontologies about human anatomy. The subject ontologies are the FMA and the Generalized Architecture for Languages, Encyclopedias and Nomenclatures for Medicine[xi] (GALEN). Both approaches use a combination of lexical and structural matching techniques. One of them additionally employs an external resource (the Unified Medical Lexicon UMLS[xii]) to obtain domain knowledge. In their work the authors point to the fact that medical ontologies contain implicit relationships, especially in the multi-word concept names that can be exploited to discover more correspondences.

The linguistic-based ontology alignment approach, which is described in the next section, builds on this finding and investigates further methods to discover the implicit information observed in concept labels of the medical ontologies. Furthermore, domain-independent ontology alignment methods are discussed by Kalfoglou and Schorlemmer (2005), Doan et al. (2003), Bruijn et al. (2006), Rahm and Bernstein (2001) and Noy (2004). We adapted techniques from all these approaches for the linguistic-based, corpus-based, and dialogue-based approach as discussed in the following.

2.1. Linguistic-based Ontology Alignment

Drawing upon our experience with the medical ontologies throughout the MEDICO project, we have identified some of the common characteristics which are relevant for the alignment process. These can be summarized as follows:

1. They are very large models.
2. They have extensive *is-a* hierarchies up to ten thousands of classes, which are organized according to different views.
3. They have complex relationships, in which classes are connected by a number of different relations.
4. Their terminologies are rather stable (especially for anatomy) meaning that they should not differ too much in the different models.
5. The modeling principles for them are well defined and documented.

Both these observations and the fact that most medical ontologies are linguistically rich suggest that linguistic-based processing of ontology concept labels (and possibly also relations) can support the alignment process. The FMA ontology, for example, contains concept names as long as *'Anastomotic branch of right anterior inferior cerebellar artery with right superior cerebellar artery'*. The linguistic processing assumes that such long multi-word terms are usually rich with implicit semantic relations (e.g., equivalences) which can be exploited to identify additional alignments.

We argue that these relations can be made explicit by observing common patterns in the multi-word terms that are typical for the concept labels in the medical ontologies. Transformation grammars[xiii] can help to detect the variants of the ontology concept labels. In other words, with the help of rules, the concept labels can be transformed into semantically equivalent but syntactically different word forms.

There some naming conventions for the complex labels of the FMA concepts. For example, the order of adjectives in the term *'Left fifth intercostal space'* is based on the rationale that the noun in the term is *'space'*; its primary descriptor is *'intercostal'*, further specified by a sequence of numbers (enhanced by the *'laterality'* descriptor).

In a similar way, the term *'Right upper lobe'* is not the preferred name of the concept, although the FMA includes it as a *'synonym of' 'Upper lobe of right lung'* because of its common usage in radiology reports (Rosse and Mejino, 2003). This means that in this example each concept label (in most cases multi-word expressions) will terminate with a noun. Some examples of the complex FMA concept labels with their lexical categories are shown in Table 1.

Bile canalicular domain of plasmalemma of hepatocyte
(noun adjective noun preposition noun preposition noun)

Blood in aorta
(noun preposition noun)

Periventricular nucleus at the tuberal level
(adjective noun preposition determiner adjective noun)

Organ with organ cavity
(noun preposition noun noun)

Pancreatic impression on spleen
(adjective noun presposition noun)

External carotid arterial subdivision
(adjective adjective adjective noun)

Table 1: Examples of FMA concept labels (preferred names and their lexical types).

One observation here is the use of prepositions (used to convey spatial information in most cases) to indicate *location* as in *'Pancreatic impression on splee'*. The prepositions we observed in these concept labels are shown in Table 2 together with their frequencies.

Rank	**Prep.**	**Freq.**	**FMA Concept Label**
1	of	119886	Bile canalicular domain **of** plasmalemma **of** hepatocyte
2	to	3167	Branch of median nerve **to** opponens pollicis
3	for	438	Atlas **for** vertebral arterial groove
4	with	263	Organ **with** organ cavity
5	in	145	Blood **in** aorta
6	between	47	Intermetatarsal joint **between** first and second metatarsal bones
7	from	42	Inferior petrosal sinus **from** pons tributary
8	on	24	Pancreatic impression **on** spleen
9	over	19	Parietal peritoneum **over** left suprarenal gland
10	within	9	Nerve ending **within** taste bud
11	behind	6	Cutaneous branch to scalp **behind** auricle
12	by	4	Esophageal impression **by** arch of aorta
13	around	3	Nodes **around** cardia
14	at	2	Periventricular nucleus **at** the tuberal level
15	below	1	Trapezoid area **below** prostate

Table 2: Prepositions observed in the FMA with their frequencies and example concept labels.

A similar statistic can be observed for RadLex. The prepositions we observed in the concept labels are shown in Table 3, together with their frequencies. Table 4 shows the transformation grammar we wrote for parsing complex medical terms.

Rank	Prep.	Freq.	RadLex Concept Label
1	of	2180	aspiration **of** lipid
2	to	58	response **to** embolization
3	with	32	dementia **with** Lewy bodies
4	for	28	marking **for** intervention
5	in	21	carcinoma **in** situ
6	from	8	satisfactory drainage **from** catheter
7	by	6	metastasis **by** lymphatic and interstitial infiltration
8	on	5	images printed **on** paper
9	around	3	out of plane wrap **around** artifact
10	at	2	loss of signal **at** interface voxels
11	between	2	partial volume averaging **between** slices
12	within	2	refocusing of selected gradients **within** one TR interval
13	behind	0	
14	over	0	
15	below	0	

Table 3. Prepositions observed in RadLex with their frequencies and example concept labels.

ConceptLabel → NounPhrase
NounPhrase → Noun
NounPhrase → Adjective NounPhrase
NounPhrase→ NounPhrase (-) Token
NounPhrase → PrivateName Noun
PrepositionalPhrase → Preposition NounPhrase
NounPhrase → NounPhrase PrepositionalPhrase

Adjective → corneal|celiac|bifurcate|selected|printed|lymphatic…
Nounn → hepatocyte|prostate|gland|intervention|embolization…
Preposition → of|in|on|at|for|within…
PrivateName → Bochdalek|Lewy…
Token → 1|2|3|4|alpha|beta|1st|2nd|X|IV|…

Table 4. The transformation grammar used to generate semantic equivalences for the common patterns in FMA and RadLex.

For example, if we take the concept label *'Blood in aorta'* from the FMA and its lexical pattern (noun preposition noun), we can apply the transformation rule,

noun1 preposition:'in' noun2 → noun2 noun1 **(1)**

and generate a syntactic variant for this concept label that nevertheless has equivalent semantics, i.e., *'Blood in aorta'* = = *'Aorta blood'*. In RadLex, this rule transforms *'Carcinoma in situ'* to *'Situ carcinoma'*. The case with the preposition 'of' in the next transformation rule is similar.

noun1 preposition:'of' noun2 → noun2 noun1 **(2)**

applies to *'Protoplasm of lymphocyte'* to generate the syntactic variant *'Lymphocyte protoplasm'*. For RadLex the same rule generates *'Lipid aspiration'* from *'Aspiration of*

lipid'. This is profitable for at least two reasons. First, it can help resolve possible semantic ambiguities (if one variant is ambiguous, it can be replaced by the other one). Second, identified variants can be used to compare linguistic (textual) contexts of ontology concepts in corpora. This leads to the corpus-based ontology alignment aspect of our approach.

2.2. Corpus-based Ontology Alignment

The basic idea of the corpus-based alignment approach[xiv] is to compare the textual and linguistic contexts of ontology classes in large corpora. We hereby assume that ontology classes with similar meanings (originating from different ontologies) will appear in similar linguistic contexts. The linguistic context can be characterized by text characteristics and computed from texts directly. These characteristics describe the data instances (i.e., the words) and attributes (i.e., the part-of-speech tags) by applying descriptive statistical measures.

Then, we will learn statistics about words and their attributes (e.g., simple occurrence frequencies or supervised information gain statistics) and use them to infer constraints that we use to associate two terms. The association is then interpreted as a candidate mapping. Corpus-based linguistics focused not only on the distribution of words, but also on the distribution of linguistic features (i.e., part-of-speech tags) which we can derive from these words in context, i.e., features about the sentences, paragraphs and texts in which a specific word or word group occurs. Analogously, the linguistic context of an ontology class to be matched to another class can be defined as:

- the document in which it appears;
- the sentence in which it appears;
- a window of size N in which it appears.

For example, a window of size +5/-5 (including stop words) for *'Antidiuretic hormone'* would be "A syndrome of inappropriate secretion of antidiuretic hormone (SIADH) was diagnosed, and bortezomib was identified as its cause." In our approach, linguistic contexts are represented by token/word vectors, (e.g., <syndrome, of, inappropriate, secretion, of, (SIADH), was, diagnosed>), <token -5, token -4, … token +4, token +5> or the following three alternative vector representations:

Binary over set of context tokens/words (e.g., 10): < 0, 0, 0, 0, 1, 0, 1, 0, 0, 0 >

Frequency over set of context tokens/words (e.g., 10): < 1, 5, 0, 0, 6, 7, 18, 1, 0, 1 >

Frequency over set of context tokens/words (e.g., 10): < 1, 5, 0, 0, 6, 7, 18, 1, 0, 1 >

Mutual Information/InformationGain over set of context tokens/words (e.g., 10):
< 1.7, 0.5, 0, 0, 1.1, 3.5, 0.5, 1.2, 0, 1.5 >

The data sets for our corpus-based experiments consisted of context data collected from the PubMed corpus on Mantle Cell Lymphoma with 1,721 scientific abstracts. 38,853 tokens matched the simple or complex terms provided by FMA, Radlex, or Image terms (the resulting set of representative image features identified by parsing a liver image showing symptoms of lymphoma) (Figure 2):

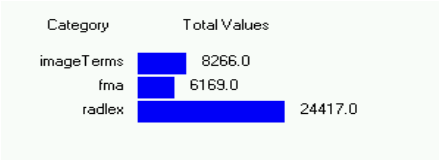


Figure 2. The distribution of term tokens that stem from the image descriptors (Image terms), FMA, and Radlex.

811 different types (terms that may represent classes) were found (FMA: 320; RadLex: 562; Image terms: 20). This means the token/type ratio for FMA is 19.28, 43.44 for RadLex, and 413.3 for the Image Terms. This means that FMA terms were not used as frequently as RadLex terms, but vary twice as much. Image terms do not show a lot of variety; a specific term (type) is used almost 414 times on average. In addition, 6,800 different context words/tokens were found. Therefore, the non-sparse vector representation for the context of each token has a dimensionality of 6,800. We then used the TnT part-of-speech tagger[xv] to annotate the POS classes (Penn Treebank Tagset[xvi]). Figure 3 shows the distribution of term tokens and the distribution of POS tags (categories).

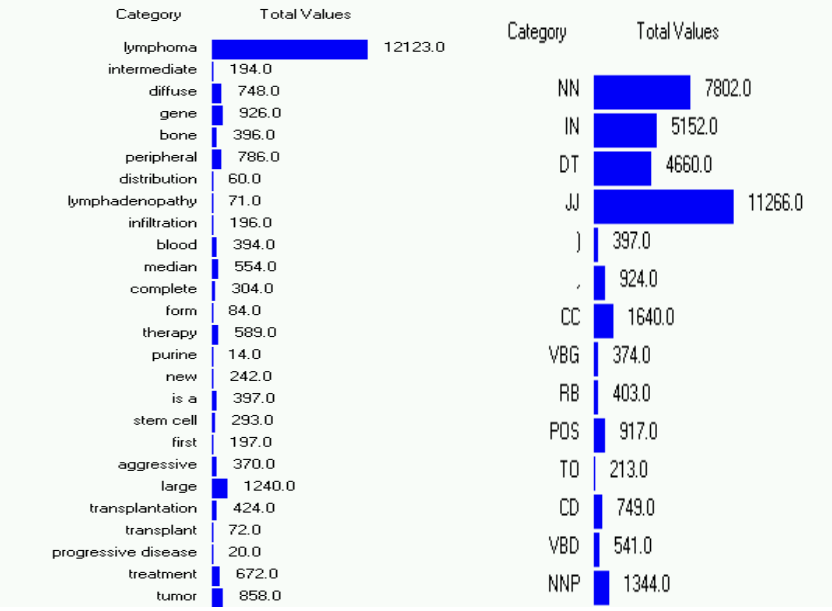


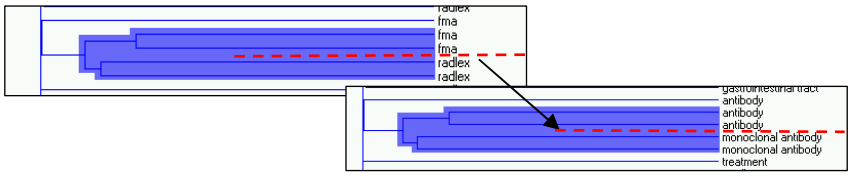
Figure 3. (Left) The distribution of term tokens that stem from the image descriptors (Image terms), FMA, and Radlex. (Right) The distribution of POS tags that stem from the image descriptors (Image terms), FMA, and Radlex.

After some experimentation with exploratory data mining methods and the medical experts, we agreed on an applicable model generation and interpretation process to be used by the medical expert. First, he generates term clusters by applying a hierarchical clustering method automatically. Then, he searches for interesting patterns in the clusters. Hierarchical clustering returns a hierarchy structure that is more informative (rather than a flat unstructured set of clusters). It does not require us to pre-specify the number of clusters or any other supervised criterion on the input data. Furthermore, it allows the expert to indicate similar meaning of corresponding ontology classes with the following procedure. He specifies a target value of interest and then searches the hierarchy for cluster boundaries.

For example, Figure 4 shows an excerpt of the cluster tree generated for the full data set #38852. The target value was set to the terms themselves (which correspond to a source value); this allows us to find new candidates for alignment because a) different terminologies

are assumed to have similar terms and b) similar terms are represented in the same cluster (or cluster boundaries) per definition. The medical expert skims the clusters (which normally refer to either FMA or RadLex) and draws attention to the shift of FMA to RadLex or RadLex to FMA as illustrated. Then, he inspects the terms of the shift (in this case *Antibody* and *Monoclonal antibody*).

Figure 4: Hierachical clustering results. The medical expert inspects the shifts from, e.g., FMA to Radlex and the corresponding terms (antibody and monoclonal antibody).



As a result, a mapping between those terms can be detected. Most importantly, this mapping was not found with a string comparison of the terms, but by clustering and interpreting the context vectors. In this way, a corpus-based method for alignment could be implemented which complements (string-based) term comparison methods and structure-based ontology alignment methods. The next step is to automize the method in order to find candidates without the expert visually mining the cluster results. Adequate measures for automatic processing are straightforward.

2.3. Dialogue-based Ontology Alignment

The ontology matching problem can be addressed by several techniques as introduced in the section on related work. Advanced incremental visualisations have also been developed (e.g., see Robertson et al., 2006) to do better than merely calculate the set of correspondences in a single shot; cognitive support frameworks for ontology mapping really involve users (Falconer et al., 2006). A dialogue-based approach could make more use of partial mappings in order to increase the usability in dialogue scenarios where the primary task is different from the schema matching task itself.

Recent work in incremental interactive schema matching stressed that users are often annoyed by false positives (Bernstein et al., 2006). This is a big problem when the user is actively engaged in the alignment process. Dialogue-based ontology alignment should provide a solution for that problem by providing a framework to elicit and interpret task-based user utterances. Task-based means that the user is not engaged in a tedious alignment dialogue where he judges proposed mappings. Instead, the doctor should use a dialogue shell to perform an intelligent image search as anticipated in MEDICO and answer only a few alignment questions if this step is not avoidable at all.

Our basic idea is as follows. Consider the methods that are required for interactive and incremental ontology mapping and evaluate the impact of dialogue-based user feedback in this process. While dialogue systems allow us to obtain user feedback on semantic mediation questions (e.g., questions regarding new semantic mediation rules), incrementally working matching systems can use the feedback as further input for alignment improvement.

In order to compute and post-process the alignments, we use the PhaseLibs library.[xvii] This platform supports custom combinations of algorithms and is written entirely in Java which allows us to directly integrate the API into the dialogue shell. In addition, the API supports

individual modules and libraries for ontology adapters, similarity measures (e.g., string-based, instance-based, or graph-based), and alignment generators.

Subsequently, we focus on interactive ontology matching and dialogue-based interaction. Rather than focussing on the effectiveness of the interactive matching approach, we describe a suitable dialogue-level integration of the matching process by example. Our interactive ontology matching approach envisions the following three stages:

1. Compute a rudimentary partial mapping by a simple string-based method;

2. Ask the user to disambiguate some of the proposed mappings;

3. Use the resulting alignments as input for more complex algorithms.

In regard to the first point, we hypothesise that the rudimentary mapping based on the concept and relation signs can be easily computed and obtained in dialogical reaction time (less than 3 seconds even for large ontologies). Second, user interactivity is provided by improving the automatically found correspondences through filtering the alignment. Concerning the third point, we employed similarity flooding since it allows for input alignments and fixpoint computation in PhaseLib's implementation following Melnik et al. (2002). The interactive semantic mediation approach is depicted in Figure 5.

In order not to annoy the user, he is only presented with the difficult cases for disambiguation feedback; thus we use the application dialogue shell basically for confirming or rejecting pre-considered alignments. The resulting alignments are then serialised as instances of an RDFS alignment format. Assuming that subsequent similarity computations successfully use the partial alignment inputs (to produce query-relevant partial alignment output), the proposed mediator can be said to be a light-weight but powerful approach to supporting incremental ontology matching and alignment generation (Sonntag, 2008).

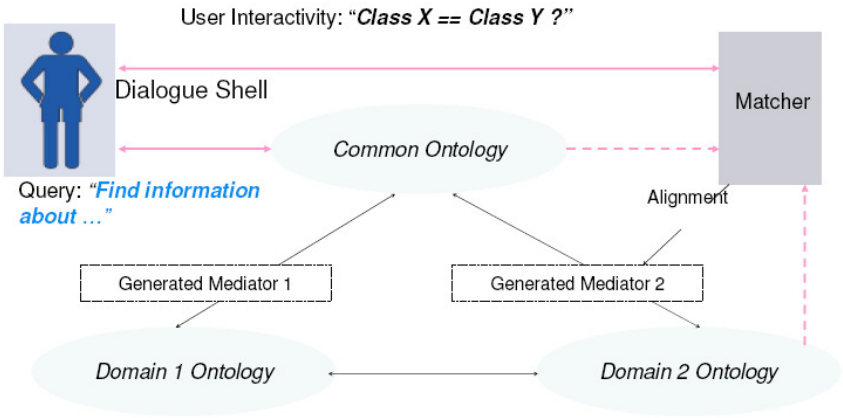


Figure 5: Dialogue-based ontology alignment approach. The user is actively involved in the alignment creation and verification process. The architecture should be particularly useful for ad hoc mappings of medical disease databases.

Figure 5 also shows the architecture of the dialogue-based ontology alignment approach. An example dialogue of the speech-based interaction between the doctor and MEDICO (we plan to implement completely in summer 2009) is shown on in Table 5:

Example Dialogue:

1. U: "Show me the internal organs: lungs, liver, then spleen and colon."

2. S: Shows corresponding patient image data according to referral record.

3. U: "This lymph node here (+ pointing gesture) is enlarged; **so** lymphadenopathy (finding Radlex Term A). Are there any comparative cases in the hospital?"

4. S: "Sorry, there are no comparative cases, shall I continue to search for similar cases?"

5. U: "Ah okay."

6. S: "Unfortunately, external databases could not be matched appropriately. Is lymphadenopathy **equivalent to** lymph node enlargement (Radlex Term B) or adenopathy (NCI Term C) or **a kind of** hematopoietic system finding (NCI superconcept)?"

7. U: **"Yes, that could be said,"**

8. U: **"a pathophysiologic process**." *(MEDICO now includes RadLex superconcepts for search.)*

9. S: "I found comparative cases in an external knowledge base according to your feedback."

Medico switches to the comparative records to help the radiologist in the differential diagnosis of the suspicious case, before the next organ (liver) is examined.

10. U: "Find similar liver lesions with the characteristics: hyper-intense and/or coarse texture ..."

11. S: Medico again displays the search results ranked by the similarity and matching of the medical terms that constrain the semantic search.

Table 5: Possible dialogue between the clinician and MEDICO

The example dialogue shows that the doctor is engaged in a task-based image retrieval dialogue. He follows the official procedure of the lymphoma finding guidelines when examining the organs. This is very important; the doctor cannot easily be asked to do anything else but search for images and complete the finding.

Fortunately, some feedback questions are allowed. With our dialogue shell (we use an upgraded version of the dialogue system for question answering on the Semantic Web we developed at DFKI, see Sonntag et al., 2007b), we try to smoothly embed the relevant question into the dialogue initiated by the MEDICO system. Most importantly, the user answers and utterances in general can be exploited for alignment judgement. Keywords, such as "so" and "equivalent to" can be interpreted to infer relations of interest, in addition to explicit user confirmations like, "Yes, that could be said."

Our datasets for a first evaluation of the three stage method as an integral part of a sensible dialogue initiative for alignments consisted of ontologies and alignment examples (manually annotated alignments for Radlex and NCI). For the first test in the medical domain, we annotated 50 alignments, 30 perfect positives and 20 perfect negatives. In the medical domain, the precision was 92% and the recall was 50% for simple string-based methods. (Corresponding concept names may differ substantially in their syntactic form.) The best matches were taken as alignment input for similarity flooding after manually confirming their validity (which simulates positive user feedback).

Our first experiments towards dialogue-based alignment generation suggest that we can use the three stage architecture as well as dialogue to do better than simply calculating the set of correspondences automatically and in a single shot. We are quite sure that in the medical domain, we cannot achieve acceptable precision and recall measurements without the expert feedback of the clinician. We are also sure that we have to obey the constraints of the doctor's task, i.e., we have to embed the alignment dialogue into the image retrieval dialogue, and, most importantly, not distract the doctor from the finding process. Furthermore, we do not start the interactive and incremental process from the refined corpus-based algorithms; it is possible to rely more on the corpus-based pre-selection by lowering the acceptance threshold of the string-based methods. But, since the dialogue-based approach is query-based, the differences cannot easily be observed. As a consequence, the linking only makes sense when the query is a "typical corpus query". According to Zipf's law, this is improbable at least for the included terms.

In future work, we are trying to provide evaluation methods to estimate the contribution of partial alignment inputs when the retrieval stage is more complex than simple name comparison, as is the case for most of our medical query patterns; user-confirmed perfect mappings can be used in simple name matching retrieval contexts with perfect precision, but this does not reflect the nature of real-world industrial requirements (in particular, where the user cannot be supposed to deliver a reliable judgement). Further, we are investigating techniques to better translate formal mapping uncertainties into appropriate dialogue-level questions for the radiologist and to address the general difficulty that users might not be able to provide helpful feedback in the course of a dialogue.

3. Ontology Population

In this section, we will deal with the semi-automatic population of ontologies from unstructured text. We will propose a methodology to semi-automatically populate the FMA medical ontology by new instances that we will derive from medical texts. We will use the query pattern mining approach explained earlier to extract relation triples from the anatomy corpus. This pattern extraction step is helped by Wikipedia-based corpora and domain ontologies; the extracted relations consist of the relation type (e.g., *known_as*, *devided_into*, or *associated_with*) and the concept instances of the relation domain and range (e.g., "vein" is associated with "artery").

The extraction of patterns corresponds to the extraction of rules from annotated text. Finally, we will apply those rules to new articles to populate the ontology. To speak from our own experience, this step cannot be achieved directly and automatically. High user input is required in order to detect and discharge the false positives.

3.1. Semi-automatic Knowledge Acquisition

Knowledge acquisition in the medical domain depends heavily on high precision. But automated ontology population provides little support for knowledge acquisition because one cannot rely on the results in terms of precision. In literature, several approaches have been proposed for, e.g., automated discovery of WordNet relations (Hearst, 1998) or discovering conceptual relations from text (Maedche and Staab, 2000).

In the medical domain, early approaches deal with the automatic knowledge acquisition from MEDLINE (Cimino and Barnett, 1993). In previous work, we evaluated potential linguistic context features for medical relation mining and designed a methodology of how to model relations and determine the parameters that distinguish relations (Vintar et al., 2003). However, all these approaches have two things in common. Either the precision values of acquired concepts and relations needed to populate a medical ontology were too low, or the task itself was too easy for the population of medical ontologies, as was the case for learning context models of medical semantic relations (Hirst and Budanitsky, 2006). Therefore, the ontology population process is time consuming and a clever semi-automatic procedure is very much in demand. To address this issue, we adapted the relation extraction approach discussed in Schutz and Buitelaar (2005) to our context in MEDICO, where the steps we took and the initial results are explained in Buitelaar et al. (2007). Having identified the statistically most relevant domain terms (i.e., ontology concepts), those about anatomy, given the domain ontology (FMA) and domain corpora (Wikipedia), we searched for relations that occur between them. For this purpose we implemented a simple algorithm that traverses each sentence, looking for the following pattern:

Noun : Verb + Preposition : Noun
(Term) (Relation) (Term)

This pattern enables us to identify possibly relevant relations between terms. The following table (Table 6) presents some early results of this work. In future work we will apply further statistical measures and linguistic heuristics to identify the most salient relations within each corpus, with an emphasis on relation identification in a more specific lymphoma corpus obtained from PubMed.

Term	**Relation (Verb+Prep.)**	**Term**
anterior	known as	anterior scalene muscle
dentate nucleus	subdivided into	anterior
muscle	situated between	anterior
body	divided into	anterior
anterior	continued over	zygomatic arch
hand	used for	anterior
artery	supplied by	medulla
artery	released if	ulnar
vein	associated with	artery
bronchopulmonary segment	supplied by	artery

Table 6: Semi-automatically extracted term-relation-term triples.

As a result, we were able to identify 1,082 non-unique relations (i.e., including syntactic variants such as *analysed_by* and *analyzed_by)*. One important requirement that comes naturally at this stage is to assess the quality and the relevance of these relations, which should be done by the clinician. Upon the clinician's approval, these relationships can be used to enrich the ontology at hand and populate the found relation instances. As discussed earlier, according to our experiences throughout the MEDICO project, the ontologies at hand do not match the requirements set by the application (see Ontology Customization under KEMM). (Most of the times there are redundancies, or important information is missing.) Semi-automatic relation extraction helps overcome this difficulty. Having identified the statistically most domain-relevant relations (about anatomy, radiology, and diseases in our case), we can customize the ontologies we use according to our domain specific needs by populating them with these relations.

More concretely, if we take the FMA, this is an ontology about the human anatomy. Similarly, RadLex provides terminology about radiology. In MEDICO, we need all this information but we need it in a specific way, i.e., in a way that relates most to the medical image semantics. As this is a very specific need, it cannot be expected that FMA, RadLex, or any other medical ontology will provide us with these relations out-of-the-box. Therefore, the relationships we have extracted from our specific domain corpora are valuable in the sense that they enable us to enhance the ontologies we use (or fragments thereof) according to our specific needs.

As an example, even though FMA comes with many relation instances, it does not contain relationships such as *'stimulated by'*. However, this relationship is present within the pattern "gastric acid *stimulated by* distention". This pattern demonstrates how terms from different ontologies (or terminologies) relate to each other specifically within the medical imaging context. Hence, including these kinds of domain specific or *custom* relationships is necessary to be able to adapt the ontologies according to our domain specific needs. In this way, we find ourselves within the portion of the FMA we use in our application. We are, however, gaining an additional radiological (and disease) perspective that comes with the relations. As important as it is to be able to extract the domain specific relations, their accuracy and relevance still need to be assessed. As our relation extraction is a semi-automatic process, it is not possible to expect no noise. Thus, correct and relevant relationships are identified as well as wrong or irrelevant ones. Sometimes, correct relationships combine with wrong terms yielding a wrong pattern altogether as in 'gene *derived from* antibody'.

The ultimate solution to avoid such noise (especially in a sensitive domain like human health) is, in our opinion, to involve the expert in the process. An effective way to involve the clinical expert in the process is to present him the relationships and their combinations with the terms (i.e., our query patterns) and ask him for feedback. In this way, the clinical expert can say whether what has been identified is correct or false. One important aspect to keep in mind is that the clinical expert is not a computer scientist. Therefore, his involvement within the process needs to be as user friendly and least technical as possible. This requirement can be fulfilled by providing him with a simple, easy-to-use and easy-to-understand interface that displays the results of the relation extraction. Upon explaining the overall objective of our task, i.e., populating (or customizing) the ontologies with what will be displayed, we can show our results using the interface. Driven by this motivation, we developed an interactive clinical query browser that displays the results of the relation extraction to the clinical expert. The next subsection gives an overview of this browser and explains its functionality.

3.2. Interactive GUI Environment for Medical Experts

The purpose of the interactive clinical query browser is to display the semi-automatically extracted domain relations and the related terms (i.e., the patterns) to the user and receive his feedback. We expect two different types of feedback. First, the expert accepts or rejects the relationships either because they are wrong or they are irrelevant. Second, he types in or dictates his general comments as free text. The relationships that he confirms will be stored and used to populate the ontology in the next step. The rest will be deleted. His free text general comments remain reference to the knowledge engineer. (In the future, however, this text may also be processed to extract further valuable information.) Figure 6 displays the views of the interface that the user sees. On the left hand side (Figure 6, (1)) all domain resources that have been used are displayed in a tree form. The first node on top 'Clinical Query Patterns' has, as of now, three children 'Foundational Model of Anatomy', 'Radiology Lexicon' and 'Image Features'. Clicking on these children nodes will display the term-

relation-term triples (i.e., the query patterns that have been identified by using the corresponding ontology or terminology).

One exception is the Image Features, where we have obtained a list of features from our partners that characterize the medical image that was automatically parsed by the image recognition algorithms. Clicking on the 'Foundational Model of Anatomy' node displays the second view as shown in Figure 6 (2). Here the patterns are displayed along with their calculated relevance scores that we explained earlier. The user has the possibility to sort this list according to any column he chooses. In the example they are sorted according to the relevance score. Each pattern can be deleted upon the user's request. The bottom pane allows the user to enter his general assessment and comments.

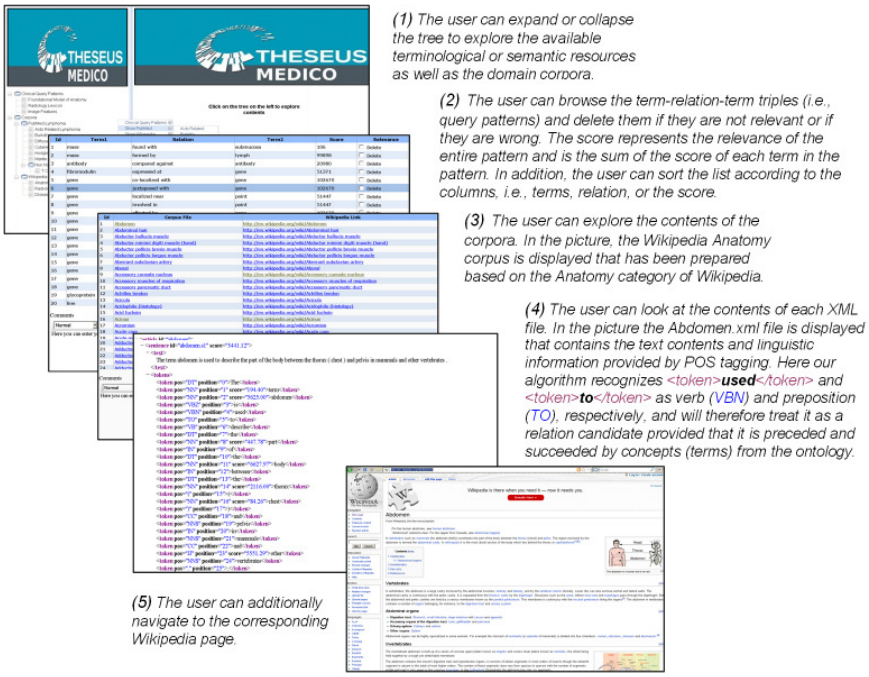


Figure 6: (1) First view that the clinician sees from the browser. The tree on the left reveals the contents available for browsing. (2) Second view of the browser that displays the query patterns as a sortable list. (3) Anatomy corpus files with links to original Wikipedia files. (4) Wikipedia *Abdomen.xml* with POS tagging. (5) Corresponding page in Wikipedia.

The 'Corpora' node of the tree has two children, which are 'Wikipedia' and the 'PubMed'. When clicked, they display the domain corpora that have been used to extract the patterns that include the relationships. Figure 6 (3) shows the Wikipedia Anatomy Corpus with the links to the corpus files and the corresponding Wikipedia pages where the files were obtained. The next example, Figure 6 (4), displays what the corpus file looks like after it has been processed to include the linguistic information that is necessary for the relation extraction algorithm. The final example in Figure 6 (5) shows the original Wikipedia page. We proposed a methodology for the population of medical ontologies; we gave the user the control over the process while automatically offering the best suggestions for the ontology population according the relation extraction step.

CONCLUSION AND FUTURE WORK

We described the three pillars of ontology treatment of the medical domain in a comprehensive case study within the THESEUS MEDICO project. These pillars are knowledge engineering, ontology mediation and alignment, and ontology population and learning. Our ontology engineering approach was constrained by the clinical knowledge requirements upon which we developed the KEMM methodology.

Concerning ontology mediation and alignment, we investigated linguistic-based, corpus-based, and dialogue-based ontology alignment. We identified linguistic features and variants that can be used to compare linguistic (textual) contexts of ontology concepts in corpora leading to the corpus-based ontology alignment aspect of our approach. In addition, we considered methods that are required for interactive and incremental ontology mapping and evaluated the impact of dialogue-based user feedback in this process.

We hypothesise that only a combination of the knowledge engineering and ontology mediation methods and rules can result in effective and efficient ontology treatment and semantic mediation. In addition, the clinician's feedback and willingness to semantically annotate images and mediation rules plays a central role, just as our capabilities to follow the official procedure of the (lymphoma) finding guidelines. In this respect, we were particularly interested in semi-automatic approaches which we not only envisioned for ontology alignment, but also for the population of ontologies. We tried to provide a semi-automatic knowledge acquisition procedure and implemented an interactive GUI environment for the medical expert. In order to ease the task of determining whether the recommended instance to be populated is correct or not, we implemented a GUI environment for the medical expert and demonstrated its interactive use by example.

In future work, we will investigate techniques to better translate formal mapping uncertainties into appropriate dialogue-level questions or suggestions displayed in a GUI for the radiologist. Furthermore, we aim to address the general difficulty that users might not be able to provide helpful feedback in the course of a dialogue or an offline GUI environment session.

A nice GUI feature to have would be the possibility to use previously found instances or classes. For example, new instances could be populated when using previously found domain or range values. In this way, a partly correct relation instance (automatically found) could be effectively re-used. This would enable the user to provide even more constructive feedback, rather than a pure reject/accept signal. This would extremely enhance the usability of the GUI tool and the effectiveness of the expert user's involvement as anticipated, particularly by the dialogical interaction scenario. In addition, the efficiency of the semi-automatic annotation approach could be improved by increasing the precision of the mappings presented to the medical expert. As experimentation shows, most time gets lost when trying to single out the false positives. Additionally, the terminologies for existing medical knowledge might change or should be expanded. Both aspects require ontology evolution, which may be addressed by an ontology learning strategy, specifically from text data about contemporary medical issues that are available in the form of the incoming patient records and new scientific articles.

REFERENCES

Bernstein, P.A., Melnik, S., & Churchill, J.E. (2006). Incremental Schema matching. In: 32nd International Conference on Very Large Data Bases, VLDB Endowment, (pp. 1167-1170).

Buitelaar P., Wennerberg P.O., & Zillner S (2008). Statistical Term Profiling for Query Pattern Mining. In: ACL BioNLP, Columbus, Ohio.

Chalupsky H. (2000). Ontomorph. A Translation System for Symbolic Knowledge. In: Proc. 17th Intl. Conf. on Principles of Knowledge Representation and Reasoning KR'2000, (pp. 471-482).

Cimino J.J, & Barnett G.O. (1993). Automatic Knowledge Acquisition from MEDLINE. In: Methods of Information in Medicine, Volume 32(2), (pp. 120-130).

de Bruijn J, Ehrig M., Feier C, Martín-Recuerda F., Scharffe F., & Weiten M. (2006). Ontology mediation, merging and aligning. In: Davies J, Studer R, Warren P (Eds.), Semantic Web Technologies: Trends and Research in Ontology-based Systems. PA: Wiley.

Doan A., Madhavan J. Domingos, P., & Halevy A (2003). Ontology matching: A machine learning approach. In: Handbook on Ontologies in Information Systems. Staab S, Studer R. (Eds.), PA: Springer-Verlag, (pp. 397-416).

Euzenat, J., & Shvaiko, P. (Ed.). (2007). Ontology matching. PA: Springer-Verlag.

Grüninger M, & Uschold M. (1996). Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 1(2), (pp. 93-155).

Hearst M.A. (1998). Automated discovery of Wordnet relations. In: Fellbaum, Ch., (Ed.), WordNet: An Electronic Lexical Database PA: MIT Press, (pp. 131-151).

Hirst A. & Budanitsky G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. In: Computational Linguistics Volume 32. PA: MIT Press, (pp. 13-47).

Johnson H.L., Cohen K. B., Baumgartner W.A. Jr., Lu Z., Bada M., Kester T., Kim H., & Hunter L. (2006). Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. In: Pac. Symp. Biocomputing, (pp. 28-39).

Kalfoglou Y. & Schorlemmer M. (2005). Ontology mapping: The state of the art. In: Semantic Interoperability and Integration, ser. Dagstuhl Seminar Proceedings, Kalfoglou Y., Schorlemmer M., Sheth A., Staab S., & Uschold M. (Eds.) no. 04391. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl.

McGuinness D., Fikes R., Rice J., & Wilder S. (2000). The Chimaera Ontology Environment. In: 17th National Conference on Artificial Intelligence (AAAI).

Madhavan J., Bernstein P., Chen K., Halevy A., & Shenoy P. (2005). Corpus-based schema matching. In: Proceedings of ICDE, (pp. 57-68).

Maedche A. & Staab S. (2000). Discovering Conceptual Relations from Text. In: (Ed.) W. Horn, Proceedings of the 14th European Conference on Artificial Intelligence. PA: IOS Press, Amsterdam.

Magnini B., Speranza M., & Girardi C. (2004). A Semantic-based Approach to Interoperability of Classification Hierarchies: Evaluation of Linguistic Techniques. In: 20th international conference on Computational Linguistics.

Falconer, S.M., Noy, N., & Storey, M.A.D. (2006). Towards Understanding the Needs of Cognitive Support for Ontology Mapping. In Shvaiko, P., Euzenat, J., Noy, N.F., Stuckenschmidt, H., Proceedings of OM-2006.

Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In: 18th International Conference on Data Engineering, (pp. 117-128).

Noy N. & Musen M. (2000). PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In: Proceedings of the National Conference on Artificial Intelligence (AAAI).

Noy N. & Musen M. (2001). Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In: Workshop on Ontologies and Information Sharing at IJCAI.

Noy N. (2004). Tools for Mapping and Merging Ontologies: In: Staab S, Studer R, (Eds.). Handbook on Ontologies, PA: Springer-Verlag (pp. 365-384).

Rahm E. and Bernstein PA (2001). A survey of approaches to automatic schema matching. In: Proceedings of VLDB, 10(4), (pp.334-350).

Robertson, G.G., Czerwinski, M.P., & Churchill, J.E (2005). Visualization of mappings between schemas. In: SIGCHI conference on Human factors in Computing Systems. PA: ACM, (pp. 431-439)

Rosse, C. & Mejino, J. L. (2003). A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. In: Journal of Biomedical Informatics 36, (pp. 478-500).

Shvaiko P. & Euzenat J. (2005). A survey of schema-based matching approaches. Journal on data semantics, Volume (4), (pp. 146-171).

Sonntag D. (2007a). Embedded Distributed Text Mining and Semantic Web Technology. In Proceedings of the NATO Advanced Study Institute Workshop on Mining Massive Data Sets for Security.

Sonntag D., Engel R., Herzog G., Pfalzgraf A, Pfleger N., Romanelli M., & Reithinger N. (2007b). SmartWeb Handheld. Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services (extended version). In: (Eds.) Thomas Huang, Anton Nijholt, Maja Pantic, Alex Plentland PA: LNAI Special Volume on Human Computing, Volume (4451) PA: Springer.

Sonntag D. (2008). Towards Dialogue-Based Interactive Semantic Mediation in the Medical Domain. In: Proceedings of the Third International Workshop on Ontology Matching (OM-2008) collocated with the 7th International Semantic Web Conference.

Starbucks, W. H. (1993). “Watch were you step!” or Indian Starbuck Amid the Perils of Academe (Rated PG). A.G. Bedeion (Ed.), Management Laureates. Volume (3), (pp. 65-110).

Vintar Š., Todorovski L., Sonntag D., & Buitelaar P. (2003). Evaluating Context Features for Medical Relation Mining. In: Workshop on Data Mining and Text Mining for Bioinformatics at the 14th European Conference on Machine Learning.

Wennerberg, P. O., Buitelaar P., & Zillner S. (2008a). Towards a Human Anatomy Data Set for Query Pattern Mining Based on Wikipedia and Domain Semantic Resources. In: Workshop on Building and Evaluating Resources for Biomedical Text Mining at LREC.

Wennerberg P., Zillner S., Moeller M., Buitelaar P., & Sintek M. (2008b). KEMM: A Knowledge Engineering Methodology in the Medical Domain. In: 5th International Conference on Formal Ontology in Information Systems (FOIS), PA: IOS Press

[i] http://sig.biostr.washington.edu/projects/fm/FME/index.html

[ii] http://www.rsna.org/radlex

[iii] http://nciterms.nci.nih.gov/NCIBrowser/Connect.do?dictionary=NCI_Thesaurus&bookmarktag=1

[iv] http://www.obofoundry.org

[v] http://www.bioontology.org/ncbo/faces/index.xhtml

[vi] http://www.ncbi.nlm.nih.gov/pubmed

[vii] http://www.obofoundry.org/cgi-bin/detail.cgi?id=chebi

[viii] http://www.obofoundry.org/cgi-bin/detail.cgi?id=cell

[ix] http://www.obofoundry.org/cgi-bin/detail.cgi?id=brenda

[x] http://lucene.apache.org/java/docs

[xi] http://www.opengalen.org

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

[xiii] We use the term transformation grammar here to imply that given a set of rules, multi-word expressions can be transformed into syntactic variants that nevertheless preserve their semantics. The concept of transformation grammar was first introduced by Noam Chomsky, where the focus was on obtaining passive sentences from the active ones.

[xiv] Actually, related work on corpus-based methods for ontology alignment does not exist in literature, at least not under this name. In the ontology matching community (see Euzenat and Shvaiko, 2007, p. 65) using statistics of text corpora would best correspond to extensional matching techniques, where data analysis in the form of the frequency distributions is used. Sonntag (2007a) demonstrated the embedding of linguistic-based approaches for instance matching, such as matching the canonical word representations, into schema matching approaches.

[xv] http://www.coli.uni-saarland.de/~thorsten/tnt

[xvi] http://www.cis.upenn.edu/~treebank

[xvii] http://phaselibs.opendfki.de

APPENDIX

Acknowledgements: This research has been supported in part by the THESEUS Programme in the CTC WP4 and the MEDICO use case, both of which are funded by the German Federal Ministry of Economics and Technology (01MQ07016). The responsibility for this publication lies with the authors. Special thanks goes to our clinical partner Dr. Alexander Cavallaro, University Hospital Erlangen.