Applications of an Ontology Engineering Methodology Accessing Linked Data for Dialogue-Based Medical Image Retrieval

Daniel Sonntag, Pinar Wennerberg, Sonja Zillner

DFKI - German Research Center for AI, Siemens AG
daniel.sonntag@dfki.de
pinar.wennerberg.ext@siemens.com
sonja.zillner@siemens.com

Abstract

This paper examines first ideas on the applicability of Linked Data, in particular a subset of the Linked Open Drug Data (LODD), to connect radiology, human anatomy, and drug information for improved medical image annotation and subsequent search. One outcome of our ontology engineering methodology is the semi-automatic alignment between radiology-related OWL ontologies (FMA and RadLex). These can be used to provide new connections in the medicine-related linked data cloud. A use case scenario is provided that involves a full-fletched speech dialogue system as AI application and additionally demonstrates the benefits of the approach by enabling the radiologist to query and explore related data, e.g., medical images and drugs. The diagnosis is on a special type of cancer (lymphoma).

Introduction

Clinical research and practice deal with complex and heterogeneous data that poses challenges to realizing applications such as medical image and text search. Thus, it becomes necessary to support these applications with explicit and machine-processable medical knowledge, e.g., about human anatomy, radiology, diseases, and drugs. This knowledge can be acquired from publicly available data sources such as published data sets, which are nevertheless seldom interlinked. Using the Linked Data concept (i.e., using the Web to create typed links between data from different sources) is a first step towards semantic-based data retrieval in the medical domain. However, ontology engineering methodologies and retrieval environments of a much more complex nature are needed to bridge the gap between the heterogeneous data sources in a realistic medical application environment.

Our experiences throughout the THESEUS Medico research project (which focuses on semantic medical image search) have shown us that several types of knowledge are relevant for the annotation of the images. In other words, when radiologists examine their patients' medical images they would additionally like to know if previous diagnoses exist, if there has been a change in the case, and what kind of medication and treatment plan is foreseen. This requires the medical images to be annotated accordingly so that the radiologists can obtain all the necessary information starting with the image. Moreover, the image information needs to be interlinked in a coherent way to enable the radiologist to trace the dependencies between observations.

Linked Data has the potential to provide easier access to significantly growing, publicly available related data sets, such as those in the healthcare domain: in combination with ontology matching and speech dialogue (as AI systems) this should allow medical experts to explore the interrelated data more conveniently and efficiently without having to switch between many different applications, data sources, and user interfaces. Since Linked Data is essentially based on the RDF representation format, the data can be linked using new RDF/OWL links that become available during an advanced medical engineering process that we adapted to Linked Data. We use, e.g., the Linking Open Drug Data (http://esw.w3.org/topic/HCLSIG/LODD) of W3C's Semantic Web for Health Care and Sciences Interest Group[1] which provides publicly available data sets that include information about drugs, diseases, clinical trials, genes etc. and some interlinking.

In this paper we investigate how medical linked data sets can be used to identify interrelations that are relevant for annotating and searching medical images while using a natural speech dialogue and SPARQL. In particular, we consider the following use case: A radiologist examines the CT scan of a lymphoma patient. Lymphoma is a type of cancer that affects the lymph nodes. He or she observes that the lymph node has shrunken and he or she would like to be able to determine whether this is due to the patient's medication. In this case, the radiologist would need to know whether the patient has been undergoing a new medical therapy and if so, the radiologist would additionally like to know everything related to the drug, for example, its (side) effects, other patients using it, etc.

[1] http://esw.w3c.org/topic/HLSIG/LODD

This paper is structured as follows. The next section presents some background and an ontology engineering methodology that this work is based upon. This is followed by the detailed description of the use case in a clinical setting or application scenario. Afterwards, related LODD data sets and our interlinking approach are presented. After the discussion about the evaluation and the integration of this data into a speech-based AI interface, the paper concludes with final remarks and the outlook.

Background and Ontology Engineering Methodology

Some current approaches to information integration, i.e., to interlinking related data, concentrate on various techniques ranging from simple string matching algorithms to more complicated graph based AI approaches. More recently, there have been intensive activities around *ontology matching* (also called *ontology alignment* or *ontology mediation*), which is a special case of semantic integration that concerns the semi-automatic discovery of semantically equivalent concepts (sometimes also relations) across two or more ontologies (Euzenat and Shvaiko, 2007). In the biomedical domain, ontology matching is being applied to the ontologies registered in the Open Biomedical Ontologies (OBO)[2] framework or in the National Center for Biomedical Ontology (NCBO) BioPortal[3].

LinkedLifeData[4] is another project that provides a platform for semantic data integration based on RDF and efficient reasoning for resolving conflicts in the data (achieved by syndicating heterogeneous biomedical knowledge in a common data model).

We have specific information needs that must be satisfied by the information sources (these should be expressed by query patterns defined over a set of ontologies). In medical imaging, a single ontology is not enough to support the required complementary image annotations from different perspectives, for example anatomy, radiology, or diseases. Ontology mediation and alignment is therefore a key aspect of the *semantic information integration task* which involves AI methods for computing equality or subsumption relations.

Integrating or interlinking related ontologies, i.e., computing these relations with the help of related AI matching techniques is an essential requirement to achieve the goals of Medico. (We used the PhaseLibs Library[5] to establish semi-automatic mappings. Variants of N-grams string matches performed best for an initial mapping.) Annotating medical images necessarily requires integrated knowledge from several resources such as diseases, radiology, and anatomy. This is in line with several other requirements that have been identified and discussed in detail in Wennerberg *et* al. (2007). Accordingly, a specific knowledge engineering methodology for the medical domain (KEMM) has been designed that covers a sequence of tasks instantiated at different stages of the process (Figure 1).

We conceive of ontology alignment as an operation on the extracted ontology modules (rather than the ontologies as a whole). The objective of the semi-automatic alignment is to obtain a coherent picture of separate but nevertheless related ontology modules. In combination with structurally heterogeneous information sources (i.e., the Web of Data / Linked Data), we speak of semantic mediation. This is because different terminologies as well as different structures and contents (which are indeed significantly different from our OWL ontologies) have to be aligned.

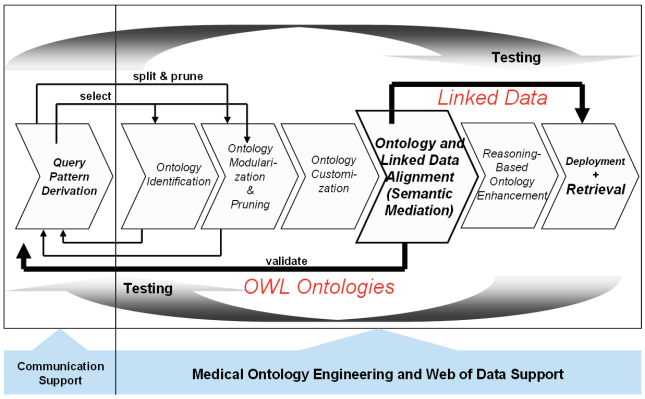

Figure 1: The KEMM medical knowledge engineering methodology workflow for semi-automatic alignment. The alignment results can be used in the Linked Data framework.

Application Scenario

Typical radiology related questions emerge from certain observations of patients' medical images. In general, the radiologists are interested in inspecting lesions, histology, changes in organs, differences between regions, etc. Example questions include, "Has an organ X changed?", "Has the lymph node disappeared?", "Are there any regional differences on this image?", "Are there any new therapies for this case?", "Are there new medicines for this case?" and so on.

As introduced earlier, we conceive of a use case in which a radiologist, who inspects the medical image of a lymphoma patient, would like to determine whether a shrinking lymph nodule is a consequence of the new medication. In this case, the radiologist starts with the image at hand (*radiology dimension*), showing an anatomical feature, i.e., the lymph nodule (*anatomical dimension*). Then, he or she moves towards diseases, i.e., lymphoma (*disease dimension*), and from there on to the medication (*drug dimension*). Clearly, when the navigation starts with the medical image, it requires the images to have been annotated previously using semantic links to connect the different dimensions.

[2]http://www.obofoundry.org
[3]http://www.bioontology.org/ncbo/faces/index.xhtml
[4]http://www.linkedlifedata.com/about
[5] http://phaselibs.opendfki.de

The already mentioned ontology matching approaches provide methods to identify the connections between the Medico ontologies that represent these dimensions. First anchors, in terms of equivalence relationships, have been established semi-automatically between the Foundational Model of Anatomy[6] (FMA) and the Radiology Lexicon[7] (RadLex), representations of human anatomy and radiology, respectively (Sonntag *et* al., 2009), (Wennerberg *et* al., 2009).

Equivalence (or relatedness) as well as sub-, superclass relationships are essential to reveal how these different but nevertheless related dimensions come together. However, they are not sufficient. To be able to gain more insight, we need additional semantic relations at a finer level of granularity. Hence, we would like to be able to know which disease affects which organ, which drug targets what disease, which anatomical features are observed in one medical image, etc. Specifically, we are interested in identifying whether a particular image modality (e.g., X-Ray) *is used for diagnosis of*, *the staging of, or the initial investigation of* a particular disease. Similarly, realizing that a particular observation on the image *is indicative of* a particular finding is relevant.

Relevant Linked Data Sets

We are interested in data sets that include information about human anatomy, radiology, diseases, and drugs as these are suggested to be most relevant for medical images by our radiology experts. In this respect, we identified the following data sets from the LODD task. The data sets below are discussed in detail by (Jentzsch *et* al, 2009).

DrugBank (Wishar *et* al., 2006) is a large repository of small molecule and biotech drugs that contains detailed information about drugs including chemical, pharmacological, and pharmaceutical data in addition to sequence, structure, and pathway information. The Linked Data DrugBank contains 1,153,000 triples and 60,300 links.[8] DrugBank defines relations such as the *drugbank:possibleDiseaseTarget* that links drugs to their related diseases. Furthermore, synonymy using *drugbank:synonym* or identity using *owl:sameAs* is included so that same / similar drugs with different names can be detected.

Diseasome (Goh *et* al., 2007) contains information about 4,300 disorders and disease genes linked by known disorder–gene associations. It also indicates the common genetic origin of many diseases. The list of disorders, disease genes, and associations between them comes from the Online Mendelian Inheritance in Man (OMIM)[9], which is a compilation of human disease genes and phenotypes. The Linked Data Diseasome contains 88,000 triples and 23,000 links[10].

DBPedia[11] is collaborative effort with the objective of extracting structured information from Wikipedia and making it available on the Web. The knowledge base consists of 274 million pieces of information (RDF triples) across several categories. The information about diseases and marketed drugs is relevant for our purposes.

As previously mentioned, to enable the navigation we start from images and move towards diseases and drugs via anatomy. Therefore we need to connect our initial resources, FMA and RadLex, with LODD data, in particular with DrugBank, Diseasome, and DBPedia. We can demonstrate this through our lymphoma use case.

Interlinking Biomedical Ontologies and Linked Data

Benefits of interlinking biomedical ontologies, in our case the FMA and the RadLex with Linked Data, become most apparent when querying for it. One main contribution of this paper is the identification of the related LODD sets for the radiology and image annotation use case. Appropriate SPARQL queries reveal the interconnections between LODD sets and their relations to human anatomy and radiology. Figure 2 shows how these different knowledge types relate to each other. A (+) or (-) sign indicates the presence (or absence) of Linked Data concepts (or the presence of instance relations) for the selected typed link.

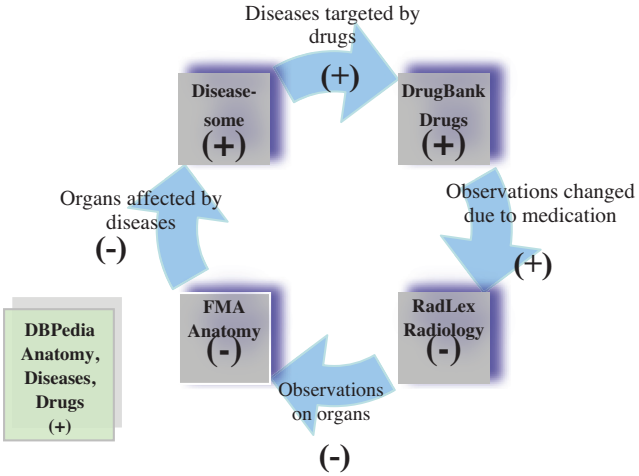

Figure 2: Conceptual workflow about how radiology, anatomy, diseases, and drugs information relate to each other.

Additionally, the DBPedia Linked Data allows for several hypothetical connections:

[6]http://sig.biostr.washington.edu/projects/fm/AboutFM.htm
[7] http://www.radlex.org
[8] http://www4.wiwiss.fu-berlin.de/drugbank
[9] http://www.ncbi.nlm.nih.gov/omim
[10] http:// www4.wiwiss.fu-berlin.de/diseasome
[11] http://dbpedia.org/About

1. **FMA (FMA/OBO, anatomy)→ DBPedia**
2. **DBPedia (drugs)→ Drugbank**
3. **DBPedia (diseases)→ Diseasesome**

These connections, however, are not discussed in the paper any further because the first "entry point", in particular the FMA/OBO mapping, is currently under development.

The SPARQL query in Figure 3 operates on the LODD data sets available from the LinkedLifeData platform, in particular those pertaining to diseases and drugs, and it lists the names of the drugs that target ***lymphoma***. Enriching the LODD data sets with additional triples of FMA and RadLex can thus enable the formulation of more complex SPARQL queries. These can finally deliver additional interlinks between drugs and organs (in **FMA** annotation) such as 'drugs' targeting diseases that occur only on a specific organ and can be detected only on a specific type of image such as CT scan (in **RadLex** annotation).

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX diseasome-diseases: <http://www4.wiwiss.fu-
berlin.de/diseasome/resource/diseases/>
PREFIX drugbank: <http://www4.wiwiss.fu-
berlin.de/drugbank/resource/drugbank/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT distinct ?drug ?drugName ?indication ?disease ?diseaseName
WHERE {
        ?drug drugbank:target ?target.
        ?drug drugbank:genericName ?drugName.
        ?drug drugbank:indication ?indication.
         ?drug drugbank:possibleDiseaseTarget ?disease.
         ?disease rdfs:label ?diseaseName.
filter(regex(?indication, "lymphoma", "i"))
}
```

Figure 3: SPARQL query that fetches drugs with indications of targeting lymphoma: FMA(-)DiseaseSome(+)DrugBank.

The SPARQL query above displays some partial interconnections among the *Diseasome* and *DrugBank* data sets. We assume that the radiologist **manually** searches for the term "lymphoma" (or also "lymph"), as this mapping from FMA anatomy to diseases simply does not exist (also cf. Figure 2, but only in the heads of the clinicians.) Next, the string "lymphoma" will be matched with all indications (?indication) of drugs (?drug) that include 'lymphoma' (or 'lymph'). Then ?drugName lists the names of the drugs (?drugName), whereas ?disease first fetches all resources from diseasome-diseases:diseases. Finally, ?diseaseName displays the disease names that are related to that drug with the given indication (i.e., lymphoma).

The combination of Linked Data sets and more structured ontological information from FMA and RadLex in OWL is particularly interesting in situations where literally looking up a concept (e.g., by using a semantic search engine) will not retrieve a target concept and, hence, also not the link to the related piece of information aimed at. In this situation, a proper ontology matching approach is demanded in order to bridge the terminology mismatch in the linked data sets. Our semi-autoamtic mapping according to the KEMM methodology should improve the situation of FMA(-)Radlex and provide the missing information for additional linkings:

- We started with existing FMA-OBO mappings available through the HCLSIG BioRDF subgroup[12] with the simple assumption that diseases or disease treatments that affect organs have name overlaps, e.g., 'large cell' (anatomical part) and 'large cell carcinoma' (a cancer type).
- Based on the output of the KEMM alignment process, a link between the RadLex concept 'Lymph node of the neck' (RadLexID: RID7695) to its semantically equivalent FMA concept (FMAID:61212) and indirectly to its super-concept 'Lymph node' (FMAID:5034) already existed. Hence, we identified a set of such initial correspondences. For example, the FMA concept 'Lymph node' is overlapped by 'Lymph node dissection' (RadLexID: RID1809), a surgical treatment that can be observed on the medical image in terms of the absence of the lymph node and the traces of the surgery.
- 'Lymph node' is then again connected to DBpedia using *rdfs:seeAlso* through the OBO-FMA mappings. Thus, FMA to DBpedia mappings provides us the access to the entire LODD linked data net. Next, through the DBpedia-DrugBank semantic link, the list of relevant drugs become available.
- Furthermore, *DrugBank* connects to *Diseasome* through the *drugbank:possibleDiseaseTarget* relation. Diseasome then lists further properties such as *diseasome:size*, that can be used to track the size of the nodules, necessary to determine the stage of the cancer.
- Eventually, we expect to be able to enhance our set of image annotations coming from our initial input resources (i.e., FMA, RadLex) with complementary semantic annotations about diseases and drugs from the LODD data set. (The result can be inspected in the resulting AI dialogue scenario).

Retrieval Architecture and Dialogue Integration

We discussed that the LODD is a valuable resource to connect the radiology dimension with the most current information about diseases and drugs via the anatomy dimension. Technically, we wrapped several Linked Data Sources into a "medical service", which count as ontology

[12] http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup/DBpedia_to_OBO_mapping

modules. These pieces of knowledge need to be interrelated within the context of the medical retrieval application, i.e., only specially predefined SPARQL queries can be used in the context of the additional Linked Data sources. Whereas the OWL ontologies can be revised and aligned in the KEMM testing and validation cycle, we focused on the deployment aspect of the mediated Linked Data sources. More precisely, this means that we deploy the semantic mediation rules in order to access the Linked Data source in the medical image retrieval scenario. Each customized ontology module represents a piece of knowledge that is necessary to realize the entire retrieval application; each Linked Data source is treated as an additional information module.

The THESEUS implementation for diverse knowledge sources (including Linked Data sources) enhances the SmartWeb implementation (Sonntag, 2009). In the current version for Linked Data Sources, the SPARQL queries have to be formulated manually. This is due to the fact that an ontological search pattern in OWL cannot be mapped to a simple SPARQL query. Often, a decomposition is needed which is not provided by our ontology matching tools (Sonntag, 2008) as a part of the *interactive semantic mediation* module. The technical intrastructure was presented in (Sonntag and Möller, 2009).

The multimodal dialogue AI system for the clinician is implemented as a native application using a special window manager for recognizing pointing gestures on a touchscreen display (Figure 4) in combination with the interpretation of natural speech requests. On the server-side, an appropriate dialogue manager is required. We use the SemVox[13] implementation which allows us to integrate the backend retrieval system while using ontology concept for message transfer. On the backend side, a dialogue manager, which uses a production rule system for input processing, provides interfaces to relevant third-party software, e.g., ASR for automatic speech recognition.

Many AI systems exist that translate natural language input into structured ontological representations or use reformulated semantic structures (see, e.g., Lopez et al., 2007). Interestingly, our rule-based natural language understanding (NLU) component delivers the concepts to be searched for in ontological form according to the medical domain ontologies. These concepts are used as the input to generate the SPARQL queries. This goes along with the use of the Semantic Web standards OWL and RDF as a common representational basis for both medical domain knowledge and annotations in the same formalism. A multimodal query pattern (speech utterance + click on a small image region) is input the for the production rule system of the dialogue manager to augment the SPARQL query. In the above dialogue (Figure 4), the radiologist's queries and, hence, the relevant Linked Data answer set depends on the specific context:

[13] http://www.semvox.de/

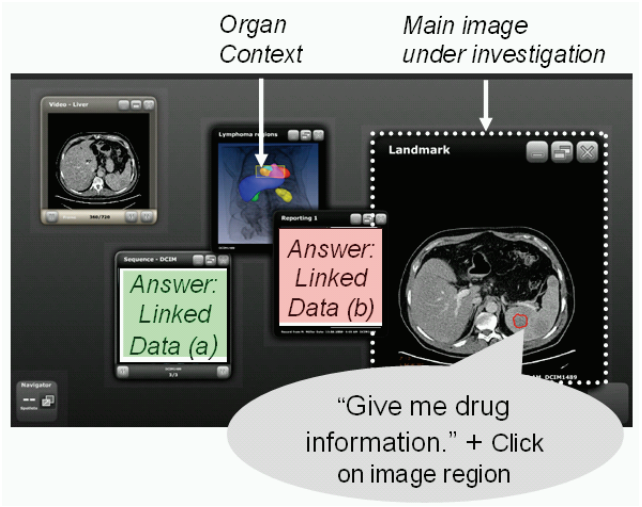

Figure 4: A radiologist's speech and touchscreen AI system installation allowing questions about image region annotations (i.e., anatomy or disease) and Linked Data in natural speech.

Given the organ context (**Answer:Linked Data (a)**), for example, head and neck, and the imaging modality, e.g., a CT scan, the radiologist comes to a differential diagnosis and eventually to a diagnosis. At this point, the radiologist can be presented with all relevant drug information that relates to his diagnosis, i.e., *lymphoma information* with the previous SPARQL query (Figure 3) that operates on the current LODD Data.

Answer:Linked Data (a)		
drugName	**Indication**	**diseaseName**
Denileukin diftitox	For treatment of cutaneous T-cell lymphoma	Severe_combined_ immunodeficiency
Denileukin diftitox	For treatment of cutaneous T-cell lymphoma	Interleukin-2 receptor, alpha chain,
Rituximab	For treatment of B-cell non-Hodgkins lymphoma (CD20 positive)	Lymphoma
Lymphoma	For treatment of B-cell non-Hodgkins lymphoma (CD20 positive)	Neutropenia
…	…	…

The only missing links here are the corresponding LODD sets for *imaging information*, i.e., LODD for anatomy as implied in figure 4 (**Answer:Linked Data (b)**).
Consequently, a SPARQL query for this purpose could in the future be formulated by combining regular expressions for indication and organ with the modality context:
filter(regex(?indication, "lymphoma", "i")),
filter(regex(?organ, "head and neck", "i")) ,
filter(regex(?modality, "CT", "i")).

Conclusions and Outlook

We can use interlinking according to the Linked Data Cloud[14] to collect additional information about a specific concept or ontology instance. This feature makes Linked Data particularly interesting when it comes to specific application domains, such as medical healthcare with complex AI dialogue systems for user interaction. However, the benefits of Linked Data can only become strong enough when the clinician can make use of the additional data in his daily task environment.

This paper discussed the benefits of semi-automatically interlinking a subset of the Linked Open Drug Data (LODD) with two semantic resources for radiology and anatomy. We argued that annotating medical images with information available from LODD can eventually improve their search and navigation through additional semantic links (in addition to our FMA / Radlex mapping, which we obtained through the KEMM ontology engineering methodology).

We explained this with a dialogue system example in the THESEUS use case Medico. The special advantage in the context of Linked Data sets is the possibility to perform automatic query expansion of the natural speech command in order to retrieve additional knowledge to semantically annotate images. These are related to semantically similar concepts in the Linked Data framework in order to overcome the limitations of current medical image systems that cannot make use of dynamic semantics.

One challenge we have encountered is the missing Linked Data sets about further biomedical resources such as radiology and anatomy. However, the OBO Foundry life science ontologies are valuable resources that can be incorporated into the LODD set. Initial work in this direction exists in form of a few hundred mappings that connect the FMA to DBPedia provided by the HCLSIG BioRDF Subgroup.[15]

Our future work concerns enhancing the LODD data set with triples from RadLex and FMA and formulating speech queries by combining regular expressions for indication and organ with the modality context. A long term goal would be to incorporate more ontologies from the OBO or NCBO portals.

Acknowledgements

This research has been supported in part by the THESEUS Program in the MEDICO Project, funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. We are thankful to our clinical partner Dr. Alexander Cavallaro from the University Hospital Erlangen, Germany, and to Kamal Najib for supporting us with the implementations.

[14] http://linkeddata.org

[15] http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup DBpedia_to_OBO_mapping

References

Euzenat, J., Shvaiko, P. 2007. *Ontology Matching*. Springer-Verlag.

Goh K.-I., Cusick M.E., Valle D., Childs B., Vidal M., Barabási A.L. 2007. *The Human disease network*. Proc. Natl. Acad. Sci. USA 104:8685-8690.

Jentzsch, A., Andersson, B., Hassanzadeh, O., Stephens, S., Bizer, C. 2009. *Enabling Tailored Therapeutics with Linked Data*. In WWW2009 workshop: Linked Data on the Web (LDOW2009).

Lopez, V., Uren, V., Motta, E., and Pasin, M. 2007. Aqualog: *An ontology-driven question answering system for organizational semantic intranets*. Web Semant., 5(2):72–105.

Sonntag, D., Wennerberg, P., Buitelaar, P., Zillner, S. 2009. *Pillars of Ontology Treatment in the Medical Domain*. In: Yannis Kalfoglou (ed.) Cases on Semantic Interoperability for Information Systems Integration: Practices and Applications. IGI Global.

Sonntag, D. 2008. *Embedded Benchmarking and Expert Authoring for Ontology Mapping and Alignment Generation*. Poster Proc. of the Fifth International Conference on Formal Ontology in Information Systems, FOIS-2008.

Sonntag, D. 2009. *Introspection and Adaptable Model Integration for Dialogue-based Question Answering*. Proc. of the Twenty-first International Joint Conferences on Artificial Intelligence (IJCAI), 2009.

Sonntag, D. and Möller, M. 2009. *Unifying Semantic Annotation and Querying in Biomedical Images Repositories* Proc. of the First International Conference on Knowledge Management and Information Sharing (KMIS), IC3K.

Wennerberg, P., Möller, M., Zillner, S. 2009. *A Linguistic Approach to Aligning Representations of Human Anatomy and Radiology*. Proc. of the International Conference on Biomedical Ontologies.

Wennerberg, P., Zillner, S., Moeller, M., Buitelaar, P., Sintek, M. 2008. *KEMM: A Knowledge Engineering Methodology in the Medical Domain*.Proc. of the 5th International Conference on Formal Ontology in Information Systems (FOIS 2008).

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J. 2006. *DrugBank: a comprehensive resource for in silico drug discovery and exploration*. Nuc. Acids Res. 1(34): D668-7.