A visual mapping tool for database interoperability: the HealthAgents case

Roman Roset[1], Miguel Lurgi[1], Madalina Croitoru[2], Bo Hu[2],
Magi Lluch i Ariet[1], Paul Lewis[2]
[1] MicroArt, Barcelona, Spain
[2] ECS, University of Southampton, UK
{rroset,mlurgi,mlluch}@microart.cat, {mc3,bh,phl}@ecs.soton.ac.uk

Abstract In this paper we present a visual mapping tool for creating mappings between a relational database schema and an ontology. Our work is motivated twofold: on one hand mapping scripts are very difficult to be written by hand by non computer experts, on the other hand practical scenarios require such mappings to be created "on-the-fly" by domain experts. We believe that our automated tool clearly illustrates the use of visual tools in the context of Semantic Web practical applications in general and eHealth applications in particular.

1 Introduction

In this paper we present a visual mapping tool for creating D2RQ [1] mappings between a relational database schema and an OWL [8] ontology. We present the practical rationale that led to the development of this tool and its functionality.

With a larger and larger number of ontologies available the need for accessing existing data stored in various relational databases has dramatically increased. This called for different languages being developed in order to support this integration via a translation between ontological and database concepts. D2RQ is one of the most widely used mapping languages due to its flexibility and compatibility easiness. It allows to specify which concepts on the ontology correspond to which concepts in the database. Once the correspondences have been established, the ontology could allow access to the database via ontology based querying languages (RDQL [2], SPARQL [4] etc.). However, mapping scripts are very difficult to be written by hand by non computer experts. In some practical scenarios such mapping should be created "on-the-fly" by domain experts who might not be necessarily accustomed to D2RQ syntax. This calls for an automated tool that allows such mappings to be developed visually. To the best of our knowledge such tool does not currently exist. This is the reason why we developed the mapping tool presented in this paper. We believe that such a tool clearly illustrates the use of visual tools in the context of Semantic Web practical applications.

The domain where we applied this tool is eHealth, in the concrete context of the HealthAgents [1] system. One of the main objectives of HealthAgents is the

[1] http://www.healthagents.net

integration of a large number of geographically distributed clinical databases of brain tumour cases. The potentially large number of acquired cases will then allow for better tumour classification using machine learning algorithms (classifiers) also distributed in a multi-agent framework. The obvious interoperability problems (different database schemas across different hospitals) was overcome by adopting a "common language": an ontology. This ontology will smooth out the integration of the schemas while also allowing for reasoning capabilities over the mediated schema. The classifiers could then select the cases that they want to train on / classify by querying the distributed databases via the ontology. This querying will require, however, a translation between the concepts of the ontology and the concepts of the database schema. If a new hospital wants to join the HealthAgents network this mapping will need to be created from scratch, usually by medical staff. At this point the need for an automated tool for creating such mapping became impetuous with clear clinical usability requirements.

Regarding related work we feel that ours was inspired by an unique challenge of mapping disparate relational databases from diverse clinical partners onto the HealthAgents ontology. Despite of this, we did some research on the literature about the current state of the art on these kind of tools.

We found that certain existing tools like KAON Reverse [3] did not handle latest standards (OWL, D2RQ) and we experienced great difficulty in using those tools for our technological needs. On the other hand, more modern tools employing the latest standards, like for example: TopBraid Composer [5]; although offering great flexibility and functionality, and in its case in particular the capability to perform mappings between ontologies, were not designed for ontology - database mapping.

For these reasons, and the lack in the literature of a tool that could be used for the purposes we wanted it for, and in the way wanted us to be used in order to facilitate the work for the clinicians using their own databases, we decided to take the challenge of designing and developing our own tool.

We will present this tool in the remainder of the paper. First, we introduce the HealthAgents system and its particular clinical requirements, in section 3 we describe the functionality of the developed mapping tool, and we conclude by presenting the ongoing work aimed at improving its usability.

2 HealthAgents

HealthAgents [6] is an agent-based, distributed decision support system (DSS) that employs clinical information, Magnetic Resonance Imaging (MRI) data, Magnetic Resonance Spectroscopy (MRS) data and genomic DNA profile information. The aim of this project is to help improve brain tumour classification by providing alternative, non invasive techniques. A predecessor project, Interpret [7], has shown that single voxel MRS data can aid in improving brain tumour classification. HealthAgents builds on top of these results and further employs multi voxel MRS data, as well as genomic DNA micro-array information for better classification results. Moreover, HealthAgents is decentralizing the Interpret

DSS by building a distributed decision support system (d-DSS). This way, the number of cases to be studied is increased, improving classifiers accuracy.

At the moment, the data in the HealthAgents system is stored in relational databases at the various participating European clinical centers. An uniform vocabulary needed for interoperability reasons is provided by means of HADOM (HealthAgents Domain Ontology). It conceptualises the parameters of the employed techniques (MRI, MRS, DNA microarrays etc.), the clinical information (age, sex, tumor location etc.) and the known brain tumor classes compliant to WHO (World Health Organisation) [2].

The patient concept is at the center of HADOM (see Figure 1(a)). Each visit of a patient is given an unique ID to be differentiated from other EHR regarding the same person. A particular patient instance, therefore, has several associated patient records. Tissue focus defines instances of the concerned areas under two sub groups, namely Primary_Focus and Secondary_Focus. A particular focus is related to one visit of a patient via Patient_Record (see Figure 1(b)). Many medical instruments and methods have been developed to diagnose brain tumour. In HADOM, we enumerate the following approaches and define them as sub-concepts of Medical_Control: Biopsy, HRMAS, Magnetic_Resonance, and Microarrays.

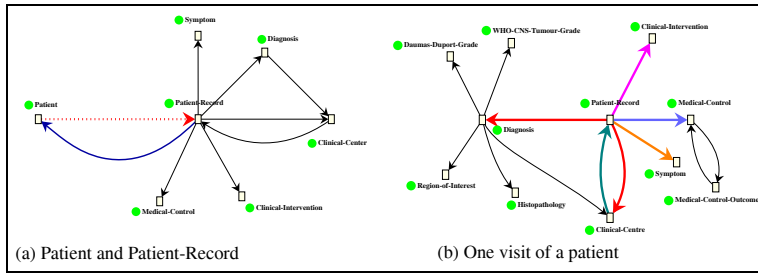

Figure 1. Conceptual view of HealthAgents HADOM

The HADOM ontology provides the basic terminology for the HealthAgents database schema and allows for interoperability at the terminological level. This is illustrated in Figure 2.

At the middle of the development we found ourselves in the necessity of integrating a number of databases (distributed over the nodes belonging to the HealthAgents network), each one with its own schema and different from each other. Obviously, the ontology framework provided by the system itself was the straightforward way to do this, however, it was a tedious job to map all the data stored in each database to the ontology, and so was born the idea of having a visual tool for doing this job in an easier way.

[2] Available from Harvard Medical School at: http://neurosurgery.mgh.harvard.edu/newwhobt.htm

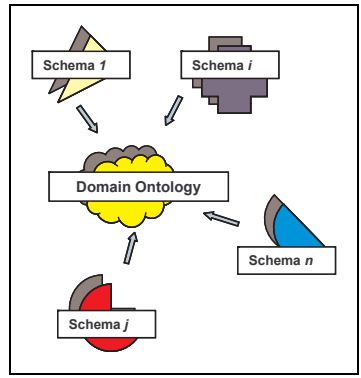

Figure 2. Ontological interoperability of HealthAgents database schema

3 Approach

The HealthAgents Mapping Tool is a software application developed for mapping between a given relational database and a given OWL ontology. Based on the idea of automating the mapping process between an ontology (concepts and properties) and a relational database schema, we designed a tool inspired by the drag and drop paradigm. This tool allows the user to relate concepts on a given ontology to entities present within a relational database with the final goal of obtaining a mapping description, using the D2RQ language for its representation.

The D2RQ framework contains a mapping language for treating non-RDF relational databases as virtual RDF graphs, and a platform that enables applications to access these graphs through the Jena and Sesame APIs, as well as over the Web via the SPARQL Protocol and as Linked Data.

The D2RQ language offers great flexibility from the point of view of mapping relationships, since it allows for a great range of properties and relations between concepts to be represented. The full specifications of the D2RQ language as well as the description of the entire platform are described in the D2RQ specifications web page [3]; where all the concepts, properties, and relationships that can be represented within a mapping description created using this language are presented.

The mapping description introduced above will allow any given user, who does not need to know the organisational schema of the database, to query the database via the SPARQL language on the ontology.

Basically, the tool facilitates the production of a mapping description between an OWL ontology and a relational database. This need for automated mapping tools is even greater in a Semantic Web era when ontologies are commonly used for interoperability and most of the data still resides in local database schemas.

[3] Available at: http://www4.wiwiss.fu-berlin.de/bizer/d2rq/spec/

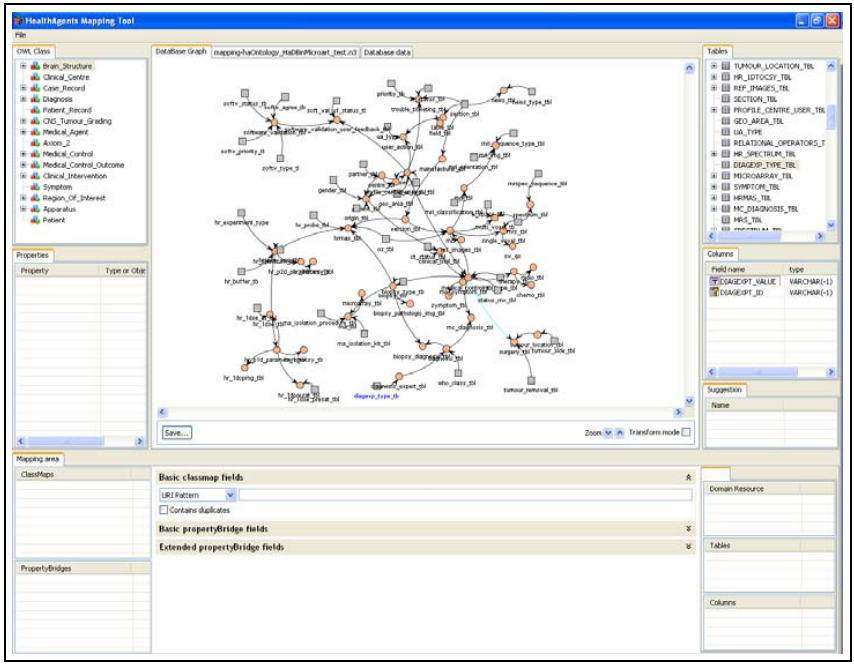

Figure 3. The HealthAgents Mapping Tool.

In the following, we present the functionality of the tool we developed and its usage in a practical setting. This will allow us to illustrate the approach taken for its design, and will also convey the full extent of the tool's capabilities. In doing this, we will start by roughly describe a typical workflow of an user working with the application.

The user is presented, on one hand, with the possibility of loading an OWL ontology that will be visualised through the built in interface. At the same time, the user can load a relational database schema specified either by an XML file or by access to the location of the actual database server. The database schema will be visually available through the interface provided by the application.

Once the ontology specification and the database schema are loaded into the application, the workflow begins by presenting the user with a directed graph that shows the entities (nodes) within the database and the relationships (e.g. foreign keys etc.) amongst them. Apart from this, a series of windows appear which are used to specify the concepts and entities to be related in a drag and drop manner. Figure 3 shows a screenshot of the application with an ontology and a database loaded. At the center of the window, the graph representing the database schema is displayed.

For the sake of clarity, in order to improve the application usability, the workspace is divided into four different areas used to present the different type of information involved in the mapping process:

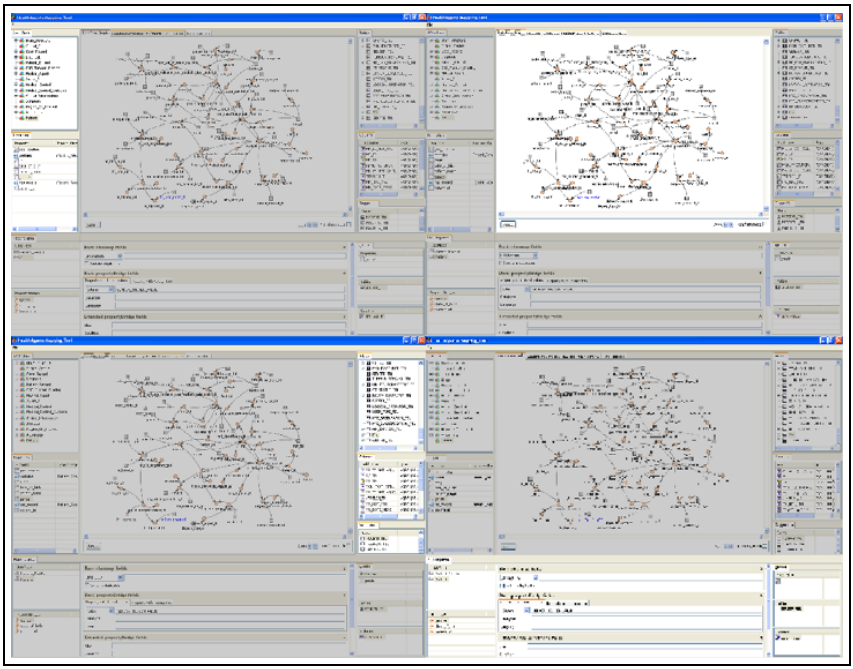

Figure 4. HealthAgents Mapping Tool spaces. Top left: the ontology area; top right: the visualisation area; bottom left: the database area; and bottom right: the mapping area

- **The ontology area:** Shows, in two windows, the concepts available within the ontology, and the attributes of the currently selected concept;
- **The visualisation area:** Apart from the graph mentioned above it presents two more tabs, displaying: the D2RQ file being generated by the mapping process, and a table presenting the data available on the database for the selected entities over the schema;
- **The database area:** Shows the schema of the specified database (tables and their fields); also makes available a window with suggestions of database schema tables for the mapping of the currently selected ontology concept;
- **The mapping area:** Displays the D2RQ specifications and the way of presenting the information of the mapping. In this space all the D2RQ specifications can be filled in to obtain a complete mapping description. This area

also fosters two subspaces: one on the side of the ontology, and the other on the side of the database; where the ontological concepts and properties, and the database tables and fields, that are being related within the current mapping description, are respectively presented.

For illustrative purposes figure 4 shows the different organisational areas used for the presentation of the information within the application. Within the figure: the top left screenshot shows the ontology area, the top right screenshot shows the visualisation area, the database area is presented on the bottom left screenshot, and finally, the bottom right screenshot presents the mapping area.

Given the intuitive character of the application's interface, the only thing to be done in order to relate a concept in the ontology (or any of its attributes) to an entity within the database is to select the desired object, drag it to the correct window within the mapping area, and do the same for the corresponding object in the database. Figure 5 shows how easily the mapping process is carried out, with a few concepts already mapped and an ontology concept being dragged to the mapping area to relate it with its counterpart within the database.

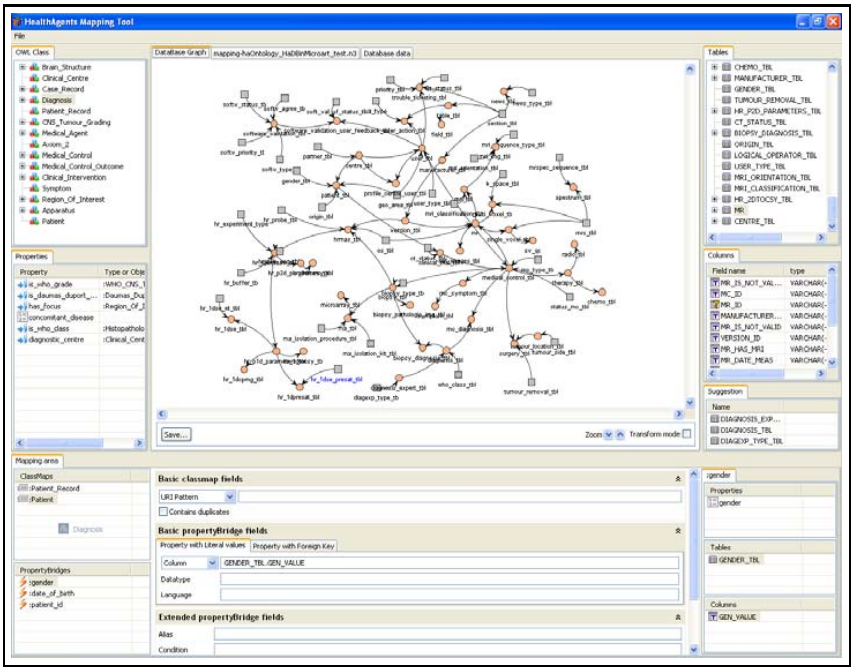

Figure 5. The mapping process. Dragging an ontology concept to the mapping area.

In order to achieve the completion of the mapping description the user must repeat the action described above for each one of the concepts on the ontology that need to be mapped. At any time during the development of the mapping, the user is allowed to visualise the final D2RQ file, which stores the mapping description. Another useful functionality is the possibility of querying the database, at any time, using the tab window provided for that specific purpose. This might help the user to decide, based on the information stored in the database, to which concept on the ontology any given entity within the database is related.

Although the mapping is done complying with the minimum specifications of the D2RQ declarative language, the application offers the possibility to create a complete mapping by presenting the user with a complete set of relations and attributes that can be specified for each of the concepts that are being mapped (following the D2RQ language specifications).

At the end of the mapping process carried out using the HealthAgents Mapping Tool application, the user has, within a D2RQ file, all the mapped concepts and the corresponding relationships with the original database entities, i.e. the mapping description (figure 6 shows an example of the D2RQ file). This allows to query that particular database at any given time, through the D2RQ platform; and also enables applications to access the database over the Web, using the mapping file generated and the SPARQL protocol as an intermediate layer between the database and the application wishing to access the data.

In the case of HealthAgents, accessing the information stored on a relational database, and mapping it to an ontology, is the main purpose for the utilization of this kind of application, and the methodology described above (using the mapping file and the SPARQL protocol) is the one which the HealthAgents system employs for accessing the data on the different databases connected to the HealthAgents network.

It is easy to see how this tool is useful for different types of applications and for people working in different areas, since the problem of accessing information stored within a relational database over the Web has been always present and in most of the cases has been sorted out with the use of ontologies.

One particular domain in which the HealthAgents Mapping Tool is essential is eHealth, given the large number of centers and clinicians willing to join the HealthAgents network for sharing data. The information they possess is stored in their local databases, hence they need a way to make that data available on the network. One way of achieving this, currently provided by the HealthAgents network, is mapping their own schemas into the HealthAgents ontology. The HealthAgents Mapping Tool offers the possibility to make this mapping process easy, by providing the clinicians with an intuitive tool that displays the schema of the local database along with the ontology they aim to make their data compatible with.

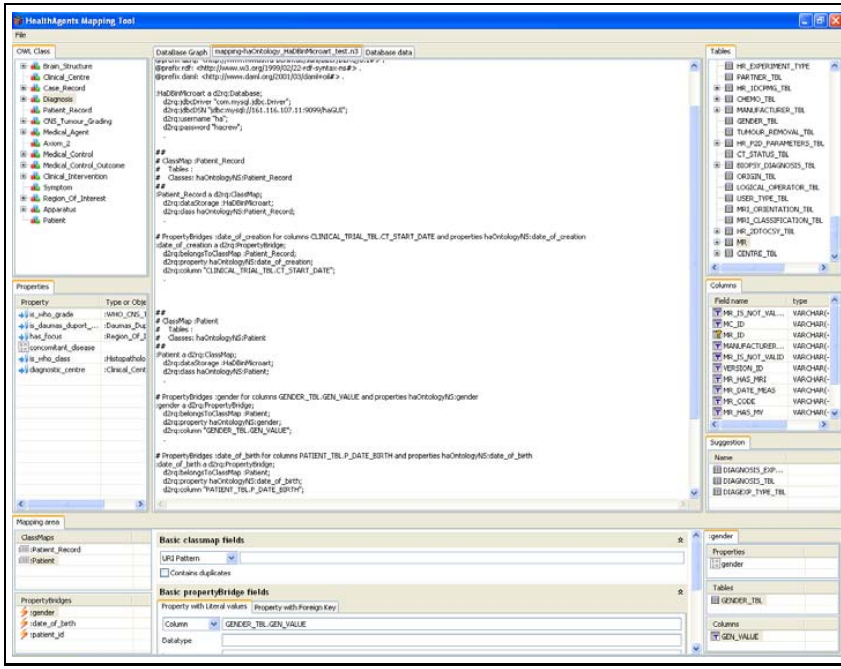

Figure 6. D2RQ file containing the mapping description.

4 Ongoing work and Drawbacks

Although the tool presented above has been successfully deployed in the context of HealthAgents, there are a number of drawbacks which we have identified. Addressing these drawbacks is current ongoing work and it would give added value for the researchers and developers using it. The following list presents some features that we think could enhance the capabilities of the tool and increase its usability:

– **The direction of the mapping:** At the moment, the mapping is done in just one direction, from the ontology to the database; in practical terms this means that the final result of the mapping process will allow the user to query the information on the database via the ontology, but will not be possible to access the information on the instances of the ontology concepts in order to use this data for storing it in the database. Looking at this fact, we think that it could be of interest to be able to perform the mapping from the database to the ontology as well, since there might be cases in which users would like to add the data in the instances of the ontology concepts to their own databases.

- **Visualisation of the ontology representation:** As mentioned in the last section, one major feature of the tool is that it shows the database schema in the form of a graph, i.e. the entities and their relationships are displayed using a directed graph on the visualisation screen, where the entities are the nodes and their relationships are the edges. But this visualisation is only available for the database schema, giving to the user a visual perspective of the database in order to facilitate the mapping process. For this purpose of facilitating the process, it could also be interesting to be able to visualize the ontology representation. There are several tools on the market that offer this functionality, and we think that it is of interest for the HealthAgents Mapping Tool to incorporate this functionality as well, allowing the user to visualize not only one side of the mapping, but both of them, for increasing the intuitiveness of the process and making it more natural.
- **Visualisation of intra mapping relationships:** When carrying out the mapping, it can be useful for the user to visualise the relationships that have been created so far, until a given state of the mapping process, between the entities within the database and the concepts on the ontology, with the objective of making easier for the user to visualise the concepts that have already been mapped, the ones that still remain to be mapped, and the relationships that exist between the objects of the two data representations. Again, as the previously presented desired features, this will facilitate and accelerate the mapping process.

5 Conclusions

In this paper we presented a visual mapping tool for creating correspondences between ontological and database concepts using the D2RQ language. The need of such tool arose in the context of the HealthAgents system where clinical partners are required to map their database schema to the system's ontology in order to support interoperability. We presented the functionality of the tool and a real case use of such system where an existing database (in Microart, Barcelona) was mapped to the developed HealthAgents ontology, HADOM. We concluded the paper with ongoing directions of work addressing different usability aspects aimed at improving the tool both from a practical view point (according to clinical feedback) but also from a research direction (ontology visualisation, ontology querying completeness).

Acknowledgements

This work is supported under the HealthAgents STREP project funded by EU Framework 6 under Grant number IST-FP6-027213.

References

1. The D2RQ Platform- Treating Non-RDF Databases as Virtual RDF Graphs. http://www4.wiwiss.fu-berlin.de/bizer/d2rq/. Last visisted 11/04/2008.

2. RDQL - A Query Language for RDF. http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/, January 2004. Last visited 11/04/2008.
3. KAON Reverse. http://kaon.semanticweb.org/alphaworld/reverse/, January 2008. Last visited 11/04/2008.
4. SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/, January 2008. Last visited 11/04/2008.
5. Top Braid Composer. http://www.topquadrant.com/topbraid/composer/index.html, May 2008. Last visited 26/05/2008.
6. C. Arús, B. Celda, S. Dasmahapatra, D. Dupplaw, H. González-Vélez, S. van Huffel, P. Lewis, M. Lluch i Ariet, M. Mier, A. Peet, and M. Robles. On the design of a web-based decision support system for brain tumour diagnosis using distributed agents. In *WI-IATW'06: 2006 IEEE/WIC/ACM Int Conf on Web Intelligence & Intelligent Agent Technology*, pages 208–211, Hong Kong, December 2006. IEEE.
7. A. R. Tate, J. Underwood, D. M. Acosta, M. . Julia-Sape, C. Majos, A. Moreno-Torres, F. A. Howe, M. van der Graaf, M. M. Lefournier, F. Murphy, A. Loosemore, C. Ladroue, P. Wesseling, J. L. Bosson, A. W. Simonetti, W. Gajewicz, J. Calvar, A. Capdevila, P. Wilkins, A. C. Bell, C. Remy, A. Heerschap, D. Watson, J. R. Griffiths, and C. Arus. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. *NMR Biomed*, 19:411–434, 2006.
8. W3C OWL Reference. OWL Web Ontology Language Reference — W3C Recommendation 10/2/2004. http://www.w3.org/TR/owl-ref/, 2004.