Using RDF and Deductive Databases for Knowledge Sharing in Healthcare

Fabiane Bizinella Nardon[1,2], Lincoln de Assis Moura Jr[1,2], Beatriz de Faria Leão[3]

[1] Escola Politécnica, University of São Paulo, Av. Dr. Luciano Gualberto, 380,
05508-900 São Paulo, SP, Brazil
fabiane.nardon@uol.com.br
[2] Atech Tecnologias Críticas, Rua Funchal, 263, 9° Andar,
04554-020 São Paulo, SP, Brazil
lincoln@atech.br
[3] Ministry of Health, Esplanada dos Ministérios, Anexo, Ala B, 1° Andar,
70058-900, Brasília, DF, Brazil
bfleao@attglobal.net

1. Introduction

Throughout a person's lifetime they may receive treatment from many different healthcare providers and each of them will store a piece of the person's medical history. Having that information available in a way that it could be easily retrieved is important not only for the patient, but also for research purposes. Each medical organization usually has its own information system and even inside a single healthcare organization, information tends to be distributed over different departmental systems, with a particular combination of software and hardware platforms. Besides that, there is a large number of different medical vocabularies that can be used to code the same health information in different levels of granularity and for different purposes. This inherently heterogeneous environment makes the task of sharing healthcare information a challenge to be met.

The representation of medical knowledge is not a trivial task, since it is necessary to use a highly expressive, platform-independent formalism that would allow several scattered peers to communicate and exchange knowledge. There is also need for a powerful query language to be able to make complex queries over complex data.

Obviously, the need for data integration and knowledge sharing does not exist only in healthcare and the Semantic Web initiative has been addressing this problem with standards like RDF (Resource Description Framework), DAML+OIL [Horrocks *at al.*, 2001] and OWL (Ontology Web Language). This paper presents our experience in using the RDF and DAML+OIL standards for data integration and knowledge sharing in healthcare. Our approach is to use a representation of the UMLS [NLM, 2003] semantic network in RDF/DAML+OIL as the basic ontology for medical concepts and a deductive database system for inferring and querying the knowledge base. As a test case, an ontology was created for the Brazilian National Health Card data interchange format, a standard for capturing and transmitting health encounter information throughout the country. Since Brazilian health providers have to integrate somehow their data to that standard, this ontology can be the starting point of a huge distributed medical knowledge base.

2. Knowledge Representation and Sharing in Healthcare

Standards for medical knowledge representation and sharing have been subject of research for many years, but none of the proposed standards reached a level of acceptance that would allow sharing health information in large scale. One of the reasons for that is the need of a very flexible and powerful formalism that would allow capturing the complexity of the medical field. The standards proposed by the Semantic Web initiative, like RDF, DAML+OIL and OWL are promising tools for addressing that problem.

There is a large number of medical vocabularies to code information in healthcare. The Unified Medical Language System (UMLS) is an effort of the U.S. National Library of Medicine aiming at facilitating the retrieval and integration of information from multiple biomedical information sources. UMLS is probably the most comprehensive ontology in healthcare, as it defines relationships among a large number of different vocabularies. In order to have a consistent basic ontology for sharing knowledge in healthcare, as part of this work we created a representation of the UMLS semantic network as a DAML+OIL ontology. The complete ontology can be found on <http://www.tridedalo.com.br/2003/07/umls/>.

The Brazilian National Health Card project is an ongoing project sponsored by the Brazilian Ministry of Health aiming at creating an infrastructure for capturing encounter information at the point of care and allowing for the construction of the national repository of clinical data. The system behind the health card stores clinical events and has a multi-level architecture (local, regional, state and federal). The system is in its pilot phase in 44 municipalities, and aims to hold information of fourteen million patients, as part of the pilot-project. In order to

explore the knowledge base created for the Brazilian National Health Card System and provide additional semantics to the data handled by it, we propose a DAML+OIL ontology for that Project. This ontology can be used not only to answer complex queries but also to help healthcare providers with legacy systems to transform their data into the Ministry of Health's format. Using deduction rules and the UMLS knowledge base, this can be accomplished. The National Health Card ontology can be found at http://www.tridedalo.com.br/2003/07/cns/.

The RDF statements created from the UMLS knowledge sources and from the National Health Card System are the basis for inferring new facts from healthcare organizations database. In this work, our approach is to use a deductive database to infer new facts and to answer complex queries, as discussed in the next section.

3. Inference and query over knowledge bases: the TRI-DEDALO system

In order to query and infer new facts from an RDF knowledge base, it is necessary a query language and an inference mechanism. Deductive Databases [Ceri *at al.*, 1990] are databases that provide all the services of a traditional Database Management System and, additionally, allow for deduction of new information using the data explicitly inserted in the database. The deduction of new facts is done by a set of deductive rules that are part of the database schema.

TRI-DEDALO (TRIples, DEduction, DAta and LOgic) is a deductive database that uses a Datalog language extension as query language. The TRI-DEDALO query language supports negation, aggregate functions, arithmetic operations, disjunction, comparison, updating and fuzzy reasoning. The main goal of the TRI-DEDALO system was to build a deductive database to be used in real world applications and the extensions added to the language are essential to achieve that goal.

The TRI-DEDALO system has also features that allow using RDF statements, or triples, as relations in the body of the rule. Also, statements can be used in the head of the rule, allowing, therefore for the inference of new statements. Figure 1 presents a few examples of TRI-DEDALO rules.

Besides the query capabilities, the TRI-DEDALO language allows to update the knowledge base, providing a syntax to express insert, update and delete operations over RDF statements.

The TRI-DEDALO system is implemented as an additional layer to any relational database. The sentences expressed in the TRI-DEDALO language are translated in a set of SQL statements that are submitted to the underlying database by the TRI-DEDALO server.

When an RDF or DAML+OIL document is loaded in a TRI-DEDALO knowledge base, its statements are translated in terms of tuples of the *statement* relation and deduction rules. For example, a DAML+OIL *sameClassAs* property would be translated as:

```
stm(x, rdf:type, y) :- stm(x, rdf:type, z),
                       stm(z, daml:sameClassAs, y).
```

Translating a DAML+OIL ontology in TRI-DEDALO rules, it is possible to use the knowledge expressed on the ontology to answer to queries using not only the explicitly inserted facts in the database, but also the derived information inferred from these facts by the deductive rules.

```
(1) query(patient?x) :- statement(subject?x,
                predicate?<http://sample.com#assignedDoctor>,
                object?<mailto:smith@samplehospital.com.br>).
(2) patientsWithCancer(patient?x) :- statement(subject?x,
                predicate?<http://sample.com#diagnosis>,
                object?<http://sample.com#cancer>).
(3) xmlns:sh=<http://www.samplehospital.com.br#>.
    patientsOlderThan18(patient?x) :-
            patients(name?x, age?y) , y > 18.
    query(name?y):-stm(s?x,p?sh:assignedDoctor,
                    o?<mailto:smith@samplehospital.com.br>) ,
            stm(s?x, p?sh:name, o?y),
            not(patientsOlderThan18(patient?y)).
(4) xmlns:voc=<http://www.avocabulary.org.br#>.
    xmlns:sh=<http://www.samplehospital.com.br#>.
    stm(x, sh:diagnosis, voc:cancer) :-
            stm(x, sh:diagnosis, y) , stm(y, voc:isa, voc:cancer).
```

Figure 1. Examples of rules using statements in TRI-DEDALO. (1) retrieves all patients that were assigned to Dr. Smith. (2) retrieves all patients with a diagnosis of cancer. (3) declares the *sh* namespace and has two rules: the first one retrieves all the patients older than 18 and the second retrieves all the patients assigned to Dr. Smith that are not older than 18. (4) derives new statements stating that a subject *x* has a diagnosis of cancer if there is a statement saying that *x* has a diagnosis *y* and *y* has a "is a" relationship with the concept cancer.

4. Conclusions

Using a combination of ontologies and the TRI-DEDALO deductive database it is possible to share knowledge in healthcare in an efficient and flexible way. The process of information integration starts with a health provider mapping their information to the UMLS and the National Health Card ontologies. Once this mapping is available, the TRI-DEDALO rules can infer information such as semantic equivalent concepts represented in different forms, hierarchy of concepts and so on, achieving this way the much needed semantic interoperability. In healthcare, a highly heterogeneous, distributed and complex domain, the possibility of sharing information can greatly improve the quality of care.

References

[Ceri *at al*., 1990] Stefano Ceri; G. Gottlob; Letizia Tanca. Logic *Programming and Databases*. Berlin: Springer-Verlag, 1990. 284p. (Surveys in Computer Science).

[Horrocks *at al.*, 2001] Ian Horrocks; et al. *DAML+OIL*. See: http://www.daml.org/2001/03/daml+oil-index.html. March 2001.

[NLM, 2003] National Library of Medicine (NLM). *Unified Medical Language System*. 14th. ed. National Library of Medicine, 2003.