Is Semantic Web technology ready for Healthcare?

Chris Wroe

BT Global Services, St Giles House, 1 Drury Lane, London, WC2B 5RS, UK
chris.wroe@bt.com

Abstract. Healthcare IT systems must manipulate semantically rich and highly structured clinical data in a distributed environment. To address this, the healthcare sector has developed standards for medical vocabulary (SNOMED-CT) and message information models (HL7 Version 3) that carry many of the features present in Semantic Web standards such as the Web Ontology Language (OWL). In this paper we examine this correspondence, and specifically present our experience in implementing SNOMED-CT. We go on to describe the fledgling use of Semantic Web technology in BT Global Services health projects and examine the obstacles to adoption of more of the Semantic Web repertoire in healthcare IT solutions.

Introduction

BT has been investigating the use of Semantic Web technology for several years in its Next Generation Web Research Group. The group led by John Davies is the coordinator of the Semantically Enabled Knowledge Technologies (SEKT) project[1], and is a core partner in the Data, Information and Process integration with Semantic Web Services (DIP) project[2]. BT's aim is to translate the use of these technologies in a research setting into benefits for our customers in divisions such as BT Global Services.

BT Global Services provides networked IT solutions for multi-site organizations. Global Services is playing a prominent role in the £6.2 billion UK National Health Service Connecting for Health (NHS CFH) National Programme for IT (NPfIT) and has won contracts worth more than £2.1 billion. One of the key services to be delivered by the National Programme is the NHS Care Records Service (NHS CRS). The NHS CRS will provide "a live, interactive patient record service accessible 24 hours a day, seven days a week, by health professionals whether they work in hospital, primary care or community services" [1]. The core of the NHS CRS is provided by the Spine, which is the name given to the national database of key information about patients' health and care. In addition more detailed patient information will be held at a local level where care is delivered. This will include records of medical conditions, medication, operations, tests, X-rays scans and other results. The scale of the National Programme has led to the local implementation of

[1] http://www.sekt-project.com/
[2] http://dip.semanticweb.org/

services being coordinated by five regional clusters. Local service providers are responsible for supporting more detailed care record information and other services at a local level within a cluster. Different activities such as radiology or hospital pharmacy will often be supported by different healthcare applications. In order to provide a care record system, these local applications will need to exchange messages containing semantically rich clinical information, and in turn summaries of this information will need to be fed to the national Spine database. BT Global Services is the national application service provider of the NHS Care Records Service providing the Spine, and the London local service provider.

In summary NHS CFH envisages a distributed system in which many diverse applications need to interoperate at a semantic level. To provide a cohesive summary of relevant clinical details for a patient, it will be necessary to aggregate information from multiple sources. These are specifically the requirements that Semantic Web technologies are being developed to address [2]. In this paper we look at the approach that is being taken to support interoperability within the National Programme, specifically a common vocabulary (SNOMED-CT), and a common information model for messaging (Health Level 7 – HL7). SNOMED-CT is a leading international medical terminology and within the National Programme, where appropriate, structured clinical information will be entered using medical terms drawn from SNOMED-CT[3]. HL7 provides standards for the definition of messages between clinical applications[4]. The National Programme is making use of the latest version 3 of HL7, which also provides an underlying information model, and specifications of how information represented using SNOMED-CT is to be conveyed within a message.

We will describe the many similarities between SNOMED-CT representation and the World Wide Web Consortium's Web Ontology Langauge (OWL)[5]. Given these similarities we also relay our experience in implementing SNOMED-CT in the hope that it informs projects implementing large scale OWL ontologies. We go on to describe the beginnings of our work that draws upon Semantic Web technology to support SNOMED-CT. Finally, we examine the barriers for more widespread adoption of Semantic Web technology within a care records system.

SNOMED-CT

SNOMED-CT is the Systemized Nomenclature of Medicine - Clinical Terms [6]. It was formed by the merger of a US medical terminology *SNOMED* with the United Kingdom medical terminology *Clinical Terms*. It aims to support the recording of clinical information using a controlled vocabulary that then enables machine interpretation whether simply for information exchange, or for decision support, aggregation and analysis. Its ongoing development is overseen by an editorial board with representatives from the College of American Pathologists and the UK National

[3] http://www.connectingforhealth.nhs.uk/technical/standards/snomed
[4] http://ww.hl7.org
[5] http://www.w3.org/2004/OWL/
[6] http://www.snomed.org

Health Service. NHS CFH has specified the use of SNOMED-CT in the Care Record Service at both a national and local level. There are several features of note.

SNOMED-CT is large with over one million terms, associated with over 400,000 concepts. SNOMED-CT is much larger than most available OWL ontologies and so poses scalability issues for OWL software tools that are only beginning to be addressed.

SNOMED-CT is concept based, in which a concept can be represented by more than one term. For example the terms '`pancreatoduodenectomy`' and '`Whipples procedure`' represent the same medical concept. Also some term strings can represent more than one concept. For example, the term '`cold`' can refer to a cold sensation concept or a common cold. It is possible to represent SNOMED-CT concepts in the OWL language as OWL classes, and different terms used to denote those classes can be represented using RDF Schema labels if required. There is nothing in SNOMED-CT equivalent to OWL instances.

Each concept is placed in a pure subsumption hierarchy within SNOMED-CT. That is, if a concept has an 'is-a' relationship with a more general concept (asthma is-a respiratory disorder), all data annotated with the more specific concept (asthma) will imply an annotation with the more general concept (respiratory disorder). The semantics of the 'is-a' relationship are equivalent to that of the OWL subclassOf axiom. For example in OWL abstract syntax:

```
SubClassOf(Asthma Respiratory_disorder)
```

Each concept may also have non taxonomic relationships with other concepts that provide more information about that concept, and may actually fully define a concept. For example, '`appendicectomy`' has an method relationship with '`excision`', a procedure site relationship with '`appendix structure`', and is fully defined. This enables applications to infer that any procedure that includes these two relationships must be an '`appendicectomy`'. The majority of these non taxonomic relationships can be regarded as existential restrictions in an OWL ontology. For example in OWL abstract syntax:

```
Class(Appendicectomy defined intersectionOf(
  Surgical_procedure
  restriction(method someValuesFrom Excision
  restriction(procedure_site someValuesFrom Appendix_structure))
```

SNOMED-CT is underpinned by a description logic (DL) based on Ontylog[3] supplied by Apelon Inc[7]. A description logic reasoner is used to check the consistency of concept definitions and classify concepts in the subsumption hierarchy. In the same way the OWL language is also underpinned by description logic. However the expressivity of the logic differs from that of Ontylog. One of the differences is the use of role grouping in SNOMED-CT [3]. Another difference is the use in SNOMED-CT of an equivalent construct to the property chain inclusion axioms planned to be a feature of OWL 1.1 [4].

SNOMED-CT is extensible at the point of data entry through the use of what is called 'post coordination'. For example, no pre-existing term exists in SNOMED-CT for 'left kidney excision', commonly referred to in medical practice. Instead, the

[7] http://www.apelon.com

terms for ‘kidney excision’ and ‘left’ exist, together with rules that specify how it is appropriate to combine them together. In an OWL ontology these would correspond with anonymous class expressions defined in terms of a number of parent classes and existential restrictions. For example in OWL abstract syntax:

```
intersectionOf(Excision
  restriction(procedure-site someValuesFrom
              intersectionOf(kidney
                       restriction(laterality someValuesFrom left))))
```

It can be seen that the move to more machine interpretable semantics with SNOMED-CT is broadly aligned with the semantics of the OWL. The examples shown in this section have been simplified for illustrative purposes and do not show the steps needed to deal with the different constructs used in the two description logics. However, those involved with the development of SNOMED-CT have made available a script to translate SNOMED-CT into OWL[8].

Health Level 7

Health Level 7 (HL7) is a standards organisation which develops message specifications to enable consistent exchange of information between healthcare applications [5]. NHS CFH have specified the use of HL7 version 3 for the messaging in the National Programme for IT. There are several features of note:

- A common reference information model (RIM) upon which all messages are based.
- For the National Programme for IT, the use of SNOMED-CT to convey the machine interpretable semantics of clinical information.
- An HL7 specific representation for the specification of the information model (not UML or OWL). Version 3 messages are commonly *implemented* using XML.

Complexity arises when clinical information can be structured using either the entities and relationships within the information model of HL7 or the concepts and relationships of SNOMED-CT. A group has been formed to work through issues in the interface between these two standards: TermInfo[9]. The ability of OWL to represent both the conceptual model of SNOMED-CT and the information model of HL7 offers an opportunity to simplify the interaction between these two standards. Rector and Marley have begun to demonstrate the utility of this approach [6].

[8] Much of the development of the mapping between SNOMED-CT and OWL and the subsequent script is the work of Kent Spackman – Scientific Director for SNOMED International.

[9] http://www.hl7.org/Special/committees/terminfo

Experience of implementing SNOMED-CT: a large ontology based terminology

As a local service provider within London, the role of BT Global Services is to integrate a collection of healthcare applications and host them. A core objective is therefore to ensure the consistent use of SNOMED-CT and HL7 by each healthcare application developer, and in some cases provide common services to be used by all applications. A key example is our development of terminology services which provide a common implementation of and access to the SNOMED-CT terminology. The initial focus is on deploying efficient term selection and browsing services so that healthcare applications can provide users with effective means of entering structured clinical information using SNOMED-CT terms. With the increased use of large ontology based terminologies to enter structured data in many domains, we expect the issues relayed here in the context of healthcare applications will be relevant to other areas.

Issues in delivering a large terminology to users

Term search: Experience using previous medical terminologies has shown that a clinician may need to enter 2-15 terms per 10 minute consultation with a patient. Considering they may have over a million terms to choose from and that these are often long, difficult to spell phrases, we must ensure the term selection process is as effective as possible. An application providing a drop down list would always be too long to use easily, but often too short to include the required term from the million available. A search box is the most straightforward solution as exemplified by leading Web search engines, but in this case we are providing search over small phrases rather than complete Web documents, and so common search strategies are not generally applicable. When performing a term search we must aim to ensure that users find all terms relevant to their search (a sensitive search strategy) and only terms relevant to their search (a specific search strategy). Increasing sensitivity decreases specificity and so we therefore have to find a balance between the two. If we search for complete phrases, relevant results are missed because of different word order. For example a search for '`strawberry allergy`' will not find a term but '`allergy to strawberries`' will. If we ask users to enter complete words to search for, it will take too long and be prone to misspelling. For example we can't expect the user to have to type in '`pancreatoduodenectomy`'. If we allow users to enter text that could appear anywhere in a word, the application often returns unexpected results. For example a user searching for '`straw all`' will intend to find '`allergy to strawberry`' but the search service will return the unexpected result '`strawberry gallbladder`'. A search for words in any order *starting with* the search strings has proven the best balance.

Focusing selection and browsing of terms to the context (subsets): SNOMED-CT has over 1 million terms as a result of its goal of being a comprehensive reference medical terminology. That is healthcare applications in many contexts designed for many purposes can all use SNOMED-CT as a common point of reference when using

medical vocabulary. However, for any one context only a fraction of the terms are relevant. To address this, SNOMED-CT has a subset mechanism. Subsets are lists of concepts or terms specified as relevant for that specific situation. For example, in an operating theatre system, a subset may be developed specific to surgical procedures. Our search services must support constrained searches within these subsets.

Hierarchies calculated using description logic reasoners (such as those in SNOMED-CT), whilst logically complete are often difficult to navigate by users. SNOMED-CT therefore also has a navigation subset mechanism in which concepts can be grouped by navigation relationships that sit outside the logical definition of those concepts. Our terminology services must therefore support applications in presenting hierarchies that follow these simpler more familiar views on the terminology.

Delivering terminology reasoning at the point of use: As already mentioned, extensibility is a central feature of SNOMED-CT through the use of post coordination (equivalent to OWL anonymous class expressions). Allowing this extensibility at the point of data entry however raises issues at every stage of the lifecycle of clinical data.

Data entry: Applications must provide an effective user interface to allow clinicians to build these expressions. The key is to present only what is sensible to construct in any one clinical context in order to reduce screen clutter with spurious options.

Data storage: Many applications expect to store a fixed length identifier for a term. Expressions can be of arbitrary length.

Data presentation: It must be possible to render the expression back into text that is familiar to clinicians. To prevent overly verbose text, this requires non trivial language generation techniques.

Data analysis: Much data analysis relies on linking specific concepts in individual patient data with more general concepts used to describe decision support rules or statistical categories. Figure 1 illustrates the architecture necessary to support post coordinated expressions. If a clinician enters a novel post coordinated expression, links must be made to the more general concepts referenced elsewhere. Therefore the application must submit these expressions to the terminology service, which in turn uses a description logic reasoner to make the subsumption links. These inferred links can then be made available back to healthcare applications and used in the execution of queries.

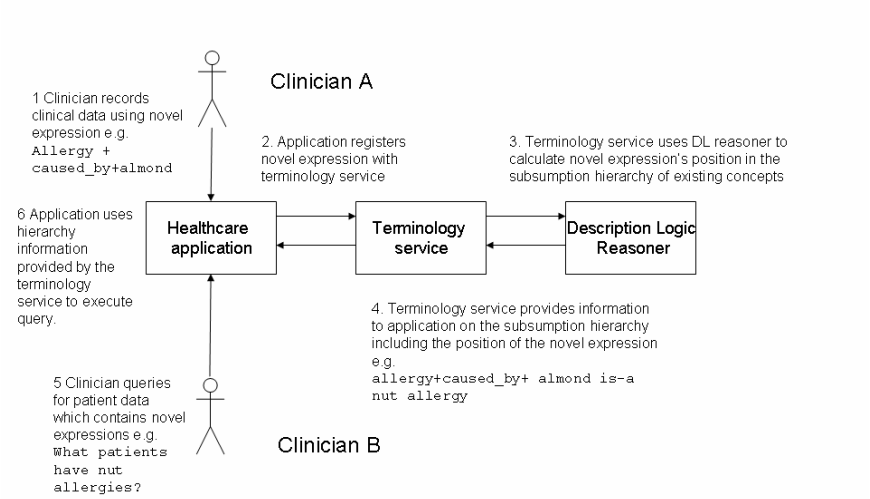

Fig. 1. The interaction between healthcare applications, a terminology service and description logic reasoner necessary to support post coordination.

Experience of using Semantic Web technology to support SNOMED-CT

We are just at the beginning of addressing the challenges of post coordination and are turning to existing Semantic Web technology to do so. The first step has been to investigate the feasibility of using description logic reasoners in the live Care Record System environment. Within BT Global Services we have developed a proof of concept system using the description logic reasoner FaCT++ from the University of Manchester[10]. This system pairs the existing terminology services used to support term selection and navigation, with the DL reasoner. As mentioned previously, SNOMED-CT developers have released a script to translate SNOMED-CT release files to OWL. This has been used as a specification to provide a more direct translation between a SNOMED-CT database and language used to interact with many DL reasoners, DIG[11]. A load process takes the SNOMED-CT release from the terminology server database, translates each statement to DIG and submits these to reasoner. To simulate the live environment, anonymous class expressions are then submitted to the reasoner as part of a subsumption query, for example `intersectionOf(allergy restriction( causative_agent someValuesFrom almond)`. That is the reasoner is asked what concepts subsume this anonymous class expression. This reflects the linking task needed between patient data and general concepts in decision support rules or statistical

[10] http://owl.man.ac.uk/factplusplus/
[11] http://dl.kr.org/dig/interface.html

categories. The result is assessed both for speed of response and validity. In this case we must ensure at least one of the results is `nut allergy`.

Initial qualitative results show that the ontology of preexisting concepts defined in SNOMED-CT can be loaded and reasoned over in an acceptable time (4 hrs on a Sunfire V210 4GB memory). Submission of test expressions as part of a subsumption query returns appropriate results in an acceptable period (<10ms for the example above). Further work is needed to assess response times with expressions of increasing complexity and also benchmark commercial alternatives to FaCT++.

Obstacles to further adoption of Semantic Web technology

Although we have found the use of OWL and associated description logic reasoners to be promising, their adoption in the eventual solution is not certain at this point. Also OWL forms only one part of Semantic Web technology. The following section describes the obstacles to wider adoption.

Lack of harmonisation between Health Informatics and Semantic Web standards: As described earlier although both OWL ontologies and SNOMED-CT are underpinned by description logic, the expressivity of the two logics is slightly different. Work needs to be done to compare the results of the two reasoning processes on the same statements to ensure the conversion from SNOMED-CT to OWL and subsequent use of OWL DL reasoners does not produce different inferences than those used in the original creation of SNOMED-CT. Only when we and our customers are confident this is the case can we use Semantic Web based DL reasoners in the Care Record System.

Obstacles in using Semantic Web technology for data representation: So far in this paper we have concentrated on Semantic Web technology purely to specify the vocabulary used to represent clinical information. With the Resource Description Framework (RDF), the Semantic Web provides a flexible graph based model to represent structured data itself with several advantages over alternative approaches including:

- a standard mechanism for the identification of resources (Universal Resource Identifier)
- a mechanism for the aggregation of data from distributed sources.
- reification which allows statements about statements. This echo's standard patient record architectures in which all entries are statements attributed to a clinical author.
- a link to the well defined semantics of OWL.

Despite this promise, the use of RDF remains at the research level within UK health informatics and is not yet being considered for implementation by suppliers in the NHS. Reasons for this include:

- **Novelty:** exposure to RDF is limited in this community

- **Alternative technology:** A relational data model is used by the majority of health care applications and clinical data warehouses. The ubiquity of this model has ensured that the tools and expertise are available with which to straightforwardly build an application. The same is not yet true for building an RDF based application or data warehouse.
- **Scalability and performance:** The flexibility of RDF comes with the downside of reduced performance. Although examples are appearing of RDF repositories containing millions of RDF statements, more evidence of performance and scalability will be needed to ensure its adoption.

Barriers in using Semantic Web technology for Web Services: At present the number of Web Services available within the National Programme has not necessitated the need for service registries or orchestration of those services. As and when the number of Web Services increases, the need for machine interpretable semantic descriptions of service functionality may then being to appear.

Conclusion

The healthcare IT sector as exemplified by the National Programme for IT appears a viable target for the adoption of Semantic Web technology. Interest in this area internationally can be gauged by the significant activity in the W3C Semantic Web for Health Care and Life Sciences Interest Group (http://www.w3.org/2001/sw/hcls/). Within BT Global Services we are beginning to explore the adoption of Semantic Web technologies specifically around the implementation of medical terminologies. However, the use of Semantic Web technology for the representation of the data, data schema and services remain the focus of research activity.

References

1. NHS Connecting for Health Implementation Guidance Team. National Programme Implementation guide v4.0, Section 3 What is the programme? (March 2006). Available at: http://www.connectingforhealth.nhs.uk/implementation

2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic Web. Scientific American 284:55 (May 2001) 28-37.

3. Spackman, K.A., Dionne, R., Mays, E., Weis, J:. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. Proc AMIA Symp. (2002) 712-6.

4. Patel-Schneider, P.F.: The OWL 1.1 Extension to the W3C OWL Web Ontology Language. Editor's Draft of 19 December 2005. http://www-db.research.bell-labs.com/user/pfps/owl/overview.html/

5. Jones, T. M., Mead, C. N.: The Architecture of Sharing. An HL7 Version 3 framework offers semantically interoperable healthcare information. Healthcare Informatics, (November 2005). Available at: http://www.healthcare-informatics.com/issues/ 2005/11_05/jones.htm

6. Marley, T., Rector, A. L.: Use of an OWL meta-model to aid message development. Current Perspectives in Healthcare Computing (2006), Conference Proceedings, Harrogate, UK, March 2006, In Press.