From Lexicon To Mammographic Ontology: Experiences and Lessons

Bo Hu, Srinandan Dasmahapatra and Nigel Shadbolt
Department of Electronics and Computer Science
University of Southampton
United Kingdom
Email: {bh, sd, nrs}@ecs.soton.ac.uk

Abstract

In this paper we describe our work on representing and constructing ontologies for breast mammography, which is part of planned set of comprehensive breast imaging and pathology ontologies designed to support the screening and diagnosis of breast cancer. We select DAML+OIL as the web-enabled ontology modelling language because of available support for description logic-based reasoning. We also discuss experiences obtained from constructing such an ontology in the hope that they can be beneficial for work on similar domains.

1 Introduction

The application of knowledge representation schemes to formalising and standardising medical protocol and terminology has had a long tradition. The emerging idea of the Semantic Web brings along with it a new era of computer-aided medical applications that enable the integration of essential information and services distributed geographically. A number of medical governing bodies have initiated the standardisation of protocols for distributed data collection to aid epidemiology and track the effectiveness of existing disease and patient management methods. Such a distributed approach increases the demands of ontologies that facilitate a shared understanding of medical vocabulary. We report on the ontology building aspect of our project that seeks to provide web-enabled services to aid the screening and diagnosis of breast cancer, and provide knowledge and image-based retrieval facilities.

Diagnosis of breast cancer normally involves multi-disciplinary meetings with experts from different medical backgrounds, e.g. radiologists, surgeons, oncologists, histologists and other clinical staff. A typical procedure for cancer assessment starts with a report from routine X-ray check or a self-report of abnormal symptoms followed by a X-ray scan. X-ray mammography is thus an obvious starting point for the knowledge modelling effort. In this paper, we present features of the mammographic ontology, the conceptual issues faced and the lessons learnt in the process.

2 Mammographic Ontology

The aim of Breast Mammographic (X-ray) Ontology (MammOnto) is to provide a commonly agreed vocabulary and formal definitions that can be used to describe the breast X-ray images, abnormal findings and medical assessments in order to facilitate knowledge sharing and reuse. As with other disciplines, we expect considerable inter- and intra-expert variability. Since it is difficult to gain access to extremely busy clinical experts, our initial effort was to construct the ontology relying on existing lexicons.

2.1 BI-RADS Vocabulary

Based on extensive field-work experience and substantial case studies, the American College of Radiology (ACR) proposed a standard for breast mammography, the Breast Imaging Reporting and Data System (BI-RADS) [1]. BI-RADS provides a comprehensive lexicon for describing mammographic findings containing: *image descriptors* (e.g. the shape of the lesion, the texture of the lesion), *lesion types* (e.g. calcification, mass), *breast cancer types* (e.g. ductal carcinoma in situ (DCIS)) and *breast cancer stages* (e.g. stage I).

We consider BI-RADS as an appropriate starting point for our MammOnto. However, the applicability of BI-RADS lexicon among British hospitals requires further validation by radiologists in the UK.

2.2 Modelling Language

Among the many available web-enabled modelling languages, we have chosen to implement the ontology in DAML+OIL [4] because it is based on description logics whose well-established model-theoretic semantics enables the definition of concept (or class) and property constructors. Moreover, DAML+OIL is supported by various ontology authoring tools which not only provide the assistance with constructing a DAML+OIL document compliant with XML standards, but also offer DL reasoning for detecting inconsistencies and making explicit knowledge implied relationships.

There are a wide variety of expressive powers provided by available DL reasoners with different combination of class and property constructors. Selecting the right expressive power is always a critical decision that one should make before going any further to construct the ontology, as: (i) there is an inevitable "trade-off" between the language expressiveness and the computational complexity of reasoning; and, (ii) expressive power has significant impact on the way in which one models the knowledge [2]. The domain knowledge of mammography requires that the selected modelling language should be able to express: universal ($\forall$) and existential ($\exists$) quantifications, qualified number restrictions ($\leq_n$, $\geq_n$), concrete property values ($\mathcal{D}$, e.g. the age of a Patient), unique properties (*f*, e.g. clinician's staff-id) and property hierarchy ($\mathcal{H}$, see Section 3.3). Therefore, together with the consideration of system availability, it gives a perfect suggestion to use $\mathcal{SHIQ}$(D) [2] as the underlying DL language for modelling and RACER [3] as the reasoner for both TBox and ABox reasoning.

2.3 Methodology

There is no fixed set of rules for building ontologies, which typically requires several stages [5]. In order to ensure that applications built on top of this ontology do not

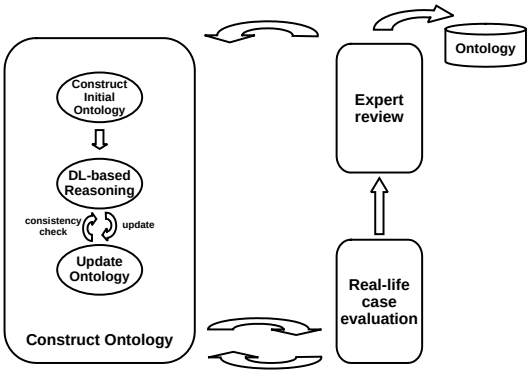

Figure 1: Local and Global Cycles of Ontology Construction

have to be substantially reconfigures, we try and maintain a modular structure dictated by their needs. In Figure 1, we illustrate a "fine tuning" approach adopted for constructing MammOnto: (1) construct the initial ontology (using BI-RADS); (2) load the ontology into a reasoner to check consistency; (3) amend, if any inconsistency is spotted, or evaluate the ontology with real-life cases and update with new knowledge, if necessary; and (4) go back to step 2 until a satisfactory ontology is obtained.

In our project, we combine the above process with a modular approach–modular in the sense that we maintain several separated tree structures, although under the same umbrella, the top-level class. A DL reasoner is responsible for the consistency checks when constructing the ontology. In order to maintain the semantic integrity of MammOnto once created, end users are not currently allowed to change the hierarchical structure or add new subclasses to existing classes.

In order to ensure the quality of MammOnto, we also set up several requirements to guide ontology development. First, anonymous existing cases are retrieved with X-ray images along with textual descriptions against which the ontology is evaluated. Second, the system should be able to support a database-like input form to encourage use among people with no experience of knowledge representation. Finally, the ontology should maintain a consistent interface which allows further extensions with no fundamental impact on applications built upon MammOnto.

3 Lessons Learned

3.1 Graininess of Modules

Despite the requirement of modularity, it was not very clear how to divide the vocabulary into several relatively independent modules. The original MammOnto was divided into four modules, namely Human, Medical-Exam, Medical-Image, and Descriptor, which proved too arbitrary for some and required refinement for others. For example, Lesion, which is initially a child of Descriptor. However, as we intend to provide some degree of inference from features appearing on X-ray images to the types of lesions, it is more appropriate to treat Lesion as a first-class citizen.

The intended image annotating application of MammOnto suggests that pure medical knowledge is not sufficient. Perceptible-Entity is introduced as the parent of all regions of interest (ROI) identifiable on the image that are suggestive of abnormalities (e.g. mass, or calcification). A bridging relation, denoted graphic-region is used to connect these subclasses of Perceptible-Entity which represent medical terms, to image

features like margin, shape, etc. which are defined as subclasses of Image-Descriptor.

Note that under the umbrella of Image-Descriptor are not only the descriptors used to represent morphological characteristics, but also those which are subject to interpretation and require medical experience and knowledge. These are listed within a sub-tree of Image-Descriptor disjoint from classes like shape.

3.2 Automated Classification

It was not evident that using DL reasoning would significantly benefit the construction of MammOnto, because of its controllable size (less than 100 classes). However, it becomes beneficial when we expect domain experts to be involved in the routine maintenance and ontology population. We also expect DL reasoning to be helpful in description retrieval for reuse as it can allocate the most general classes in an ontology which are subsumed by a particular query. For instance, one can retrieve all Round Perceptible-Entity using Query (1). Although, there is no class defined exactly as the "*round perceptible entity*", a DL reasoner is able to find the most general children of Query (1) and retrieve all their instances and the instances of their sub-classes as the answer to the query.

$$\mathsf{Perceptible\text{-}Entity} \sqcap (\forall\, \mathsf{has\text{-}shape}\, \mathsf{Round}) \tag{1}$$

Also, when extending the ontology, one can draw on subsumption to not make it mandatory for the user to explicitly specify the new class to be a sub-class of an existing class, when some of such relationships are not so obvious and available.

3.3 Hierarchical Property Structure

MammOnto was originally developed with only primary properties, i.e. properties are defined only by names. Such approach helps reduce the complexity of DL reasoning, yet introduces problems for applications based on the ontology.

Defining properties only with names results in a flat structure which is against our modular design philosophy. If each property is independent of all others, it is difficult to maintain a common interface when new properties are added and thus applications based on MammOnto will have to be suitably modified along with the ontology. For instance, in order to produce natural language reports, an application needs to handle properties such as contains and produced-by differently. When a new property involved-in is introduced, the application will have to be appropriately extended as well.

Properties are organised into different categories to tackle such issues. Currently, four primary properties are created, viz. active-action-property, passive-action-property, attribute-property and part-whole-property. Any other properties are defined as a sub-property of one of these four generic ones (e.g. $\mathsf{has\text{-}staff\text{-}id} \sqsubseteq \mathsf{attribute\text{-}property}$). Hence, applications can be developed generically based on the top-level properties and applied to their descendants (sub-properties). Note that such categories are introduced in order that there is no interference with DL reasoning. For instance, we do not specify part-whole-property to be transitive to prevent unsound propagation of properties.

3.4 Meta Classes

The transitivity of part-whole relations was encountered in other places, for example when several lesions (e.g. *calcifications*) of the same type exist on one image. A col-

lection class Calcification-Collection $\sqsubseteq$ Lesion, introduced in the initial ontology, should be interpreted as a collection of lesions rather than a type of lesion. This ambiguity can be traced to BI-RADS, which equivocates between individual *calcifications* and a collection with different size and morphological characteristics.

We found it natural to create a meta-class MetaData-Calcification with slot values for number, size, distribution, for the collection. Relations are created between *Calcification* and their meta information using role reference, e.g.

has-meta-info(Calcification, MetaData-Calcification).

4 Conclusions

MammOnto stems from research on standardisation carried out by ACR. Its applicability in British hospitals and academic environments will have to be evaluated with user-based experiments against other vocabularies. Such evaluation will be performed both passively, by modelling the retrieved anonymous cases with textual descriptions or auditing the breast cancer assessment meeting to describe the information presented during the meeting, and actively, by involving and encouraging radiologists to use the ontology.

At the moment we are trying to integrate ontologies to capture other imaging modalities, like MRI and other aspects of the multi-disciplinary procedure, notably histopathology. These introduce further challenges for modelling and integration with the MammOnto.

Acknowledgements

Research for this paper was conducted in the MIAKT project funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under the grant GR/R85150/01.

References

[1] American College of Radiology. *Breast Imaging Reporting and Data System: BI-RADS*, third edition.

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.

[3] V. Haarslev and R. Möller. *RACER User's Guide and Reference Manual Version 1.6*. University of Hamburg, Computer Science Department, July 2001.

[4] D. L. McGuinness, J. Hendler, and L.A. Stein. DAML+OIL: An Ontology Language for the Semantic Web. *IEEE Intelligent Systems*, 2002. 72–80.

[5] N. F. Noy and D. McGuinnes. Ontology Development 101: A Guide to Creating Your First Ontology. Technical Report SMI-2001-0880, School of Medical Informatics, Stanford University, USA, 2001.