MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-986

Semantic Reasoning with XML-based Biomedical Information Models

Martin J. O'Connor, Amar Das

Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA USA

Abstract

The Extensible Markup Language (XML) is increasingly being used for biomedical data exchange. The parallel growth in the use of ontologies in biomedicine presents opportunities for combining the two technologies to leverage the semantic reasoning services provided by ontology-based tools. There are currently no standardized approaches for taking XML-encoded biomedical information models and representing and reasoning with them using ontologies. To address this shortcoming, we have developed a workflow and a suite of tools for transforming XML-based information models into domain ontologies encoded using OWL. In this study, we applied semantics reasoning methods to these ontologies to automatically generate domain-level inferences. We successfully used these methods to develop semantic reasoning methods for information models in the HIV and radiological image domains.

Keywords:

Knowledge bases, Medical informatics applications

Introduction

The Extensible Markup Language (XML) has recently become an important technology in many medical domains, driven primarily by the desire for greater interoperability between biomedical software applications [1]. XML is used extensively to define *information models* that describe the structure and content of biomedical data that can be exchanged between applications. The general approach is to define the structure and content of an information model using XML Schema [2] and to publish this model to enable the production, consumption and validation of XML documents that conform to the model. Some organizations are working to define standard information models for particular domains [3]. For example, the Annotation and Image Markup (AIM) Project of the U.S. National Cancer Institute's cancer Biomedical Informatics Grid has defined an information model to describe annotations on radiological images [4]. Many custom models are produced for particular biomedical systems to support downloading or uploading of application data. Irrespective of their origin, these information models have become invaluable tools for dealing with the high degree of heterogeneity that is typical in biomedical data.

As useful as they have become, data in XML-based information models are not typically in a form that is directly suitable for reasoning. When dealing with these models, system developers must develop custom software to import model content and map it to internal application formats, where it can then be manipulated. This process is labor-intensive and time-consuming and is usually heavily customized to both the source information model and the final reasoning tasks. There is a pressing need for more principled methodologies to automate these processes. Ontologies provide a means of tackling this informatics challenge. The low-level information defined by information models can be significantly enhanced by transforming the data to domain-level content described using ontologies. Automated reasoning tasks can then be applied to the resulting domain-level information. These tasks can include classification, verification, temporal and spatial reasoning, and the generation of high-level domain abstractions that can be used directly by system users. This approach to reasoning with information model data provides an opportunity to leverage the reasoning mechanisms provided by ontology-based tools and to exploit the increasing use of ontologies in biomedicine.

Background

The Web Ontology Language (OWL; [6]) is increasingly being used in biomedical applications. Although OWL and XML share similar content storage goals, OWL provides much more powerful features both for representing semantic information about content and for reasoning with it. In combination with the OWL-based Semantic Web Rule Language (SWRL; [7]), OWL provides facilities for developing very powerful reasoning services. Many XML information models in biomedicine use standardized terms defined using ontologies. For example, the AIM information model supports the use of RadLex terms [5] when describing the anatomic structures in image observations. In general, however, beyond the use of these term references, there is no direct interoperation path between the data in XML-described information models and OWL ontologies. Hence, before reasoning services can be developed, this mapping challenge must be addressed

A variety of XML Schema to OWL mapping tools have been developed [14-15]. These tools typically provide custom mapping languages in combination with graphical user interfaces to allow users to produce OWL equivalents of XML Schema-described documents. The mappings supported by these tools

are generally low level and structural. Most could support the steps required to transform XML-described biomedical information models to their OWL equivalent. However, the more complex transformations needed to generate domain ontologies are beyond the capabilities of these tools. In biomedicine, the temporal components of these mappings can be particularly complex. An array of layered knowledge transformations are often required before the information is directly suitable for reasoning. These transformations usually demand custom solutions. The different approaches currently used to address these tasks can produce a disconnected, fragmented workflow. Integrating the various tools and methods that perform information model mapping, domain model generation, and the final reason tasks could help produce a streamlined end-to-end process that lowers the overall development effort.

A suite of tools to provide this workflow must support: (1) mapping information models to their ontolgical equivalents; (2) mapping these data to domain ontologies; and (3) developing standardized reasoning approaches for processing the resulting information. In this paper, we describe the development of such a set of tools. We outline a workflow that uses these tools to transform XML-based information models into domain ontologies encoded using OWL and then perform a variety of reasoning services on this domain knowledge. We show how we have applied these techniques to perform semantic reasoning with data contained in XML information models in the HIV and radiological image domains.

Methods

Our approach to transforming an XML-based information model to an OWL domain ontology comprises three tasks: (1) produce an OWL equivalent of an XML-based information model, (2) transform its content into instances in an OWL domain ontology, and (3) define and implement domain-level reasoning tasks that use the domain ontology. The goal is to produce an automated process that takes XML-encoded information model documents, transforms them to instances in an OWL domain ontology, and to then perform semantic reasoning with these domain-level instances.

Transforming an XML Information Model to OWL

This step requires development of an OWL ontology to represent the information in the original XML information model. The goal is to transform the XML Schema-described information model into an ontological representation that defines a semantically equivalent information model. This model must represent all the concepts in the original XML model. This transformation is performed by creating classes and properties in the OWL information model that correspond to respective components in the source information model.

We developed a tool called XMLMaster to define these transformations. XMLMaster was written as a plugin to the popular Protégé-OWL ontology development environment [8] and provides a graphical user interface that allows users to interactively define mappings between entities in an XML document and concepts in an OWL ontology. It can be used to define mappings between an XML model and an existing OWL ontology, or it can generate a new OWL ontology as the target of these mappings. We used this latter mode to create an OWL information model that corresponds to a source XML-encoded information model. These mappings were stored by XMLMaster in a mapping ontology. They contain a specification of how entities can be mapped from an XML document to instances in an OWL ontology. We then used an associated tool called XMLMapper to take the mappings and to automatically transform XML documents to OWL ontologies. As part of a workflow, XMLMapper can be used to process streams of XML documents and populate an OWL knowledge base with the resulting transformed content.

Most XML-based information models are designed to be somewhat human readable, so generally the transformations are not structurally complex. In many cases, the structure of the information model in OWL will be similar to the structure of the XML information model. However, specialized transformations are often required to deal with references to external terminologies. Such references to terms in controlled terminologies are common in biomedical information models. The mapping process must maintain these links if possible. XMLMaster supports links to these external terminologies if they are encoded using RDF or OWL. It currently does not support automatic references to terms defined to non OWL or RDF ontologies. In these cases, the original term identifiers are simply mapped unchanged, with additional annotations describing the source terminology.

Transforming an OWL Information Model to a Domain Ontology

An information model does not generally represent data in a form that is directly suitable for use in the complex reasoning typical in biomedicine. The second mapping step is generally required to transform instances in the OWL information model to instances in a specialized domain model. Domain-level reasoning tasks can then be defined using this representation. This second stage is typically far more complex than the primarily structural initial mapping process. It often requires in-depth domain knowledge and its requirements for data transformation are far more demanding. Depending on the complexity of the domain ontology, this process may require several mapping layers that operate at successively higher levels of abstraction. While several OWL-based mapping tools are available, they generally do not offer the flexibility to easily capture the full array of possible transformations required.

Instead of using one of these tools, we used OWL and its associated rule language SWRL [7] to define the mappings. SWRL provides particularly strong support for this type of knowledge transformation. Its tight integration with OWL allows it to be used to define knowledge-level mapping rules that are fully aware of OWL's complex semantics. Moreover, most OWL classifiers support SWRL, so they can be used to semantically validate these mapping rules. Once defined, the rules can be stored in an OWL ontology and later executed to transform information model instances to domain ontology instances as part of a mapping workflow.

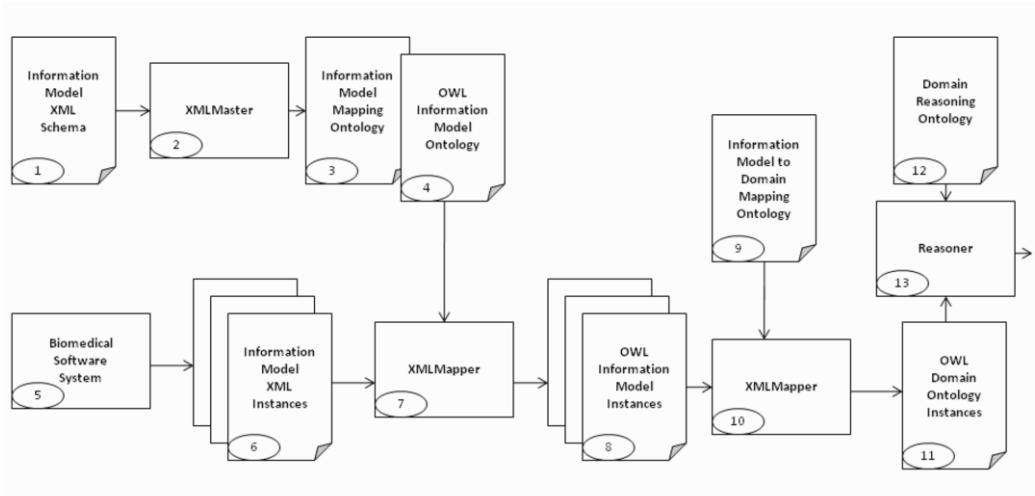

Figure 1-Outline of workflow to take XML-encoded information model instances and transform them to an OWL domain ontology and perform reasoning with them.

Reasoning Ontology and Querying

Once an XML information model is transformed to an OWL domain ontology, users can develop knowledge-level ontology-based reasoning mechanisms to work with their content. A wide array of possible reasoning tasks can then be defined. Common tasks in biomedicine include classification, and spatial and temporal reasoning. Temporal reasoning tasks are particularly central to many biomedical applications. OWL itself provides strong support for classification. It has relatively weak support for temporal and spatial reasoning, however. Fortunately, SWRL provides basic support for spatial reasoning using its core language operators. Crucially, it provides a mechanism to define custom libraries for specific types of reasoning processes. We used this extension mechanism to develop a temporal reasoning library [12].

In some cases, not all reasoning can be carried out using OWL and SWRL, and custom application methods are required. To support the necessary extraction of content from domain ontologies, we used SQWRL, a language that we developed [9]. SQWRL (Semantic Query-Enhanced Web Rule Language) is a SWRL-based query language that can be used to query OWL ontologies. SQWRL supports queries that extract information from ontologies at the knowledge level, and thus minimize the amount of custom application logic required to process this ontology-encoded information.

Defining an Automated Workflow

We defined a workflow to take an XML-based information model and perform domain-level OWL and SWRL-based reasoning with the information in the model. The numbered steps in Figure 1 outline the sequence of steps in this workflow.

We first took an XML Schema-described information model (1), and used the XMLMaster tool (2) to define mappings to an equivalent OWL information model. The XMLMaster tool also produced a mapping ontology (3) that defined how XML documents are transformed into instances of the OWL information model (4). We then produced data (5), which are encoded as document instances of the XML information model (6). These documents were fed through XMLMapper (7), which used the mapping ontology defined by XMLMaster to produce OWL information model instances (8). A separate SWRL-based mapping ontology defined how these instances were mapped to a domain ontology (9) and were used by XMLMapper to transform the instances (10) to domain ontology instances (11). Finally, a domain-level OWL reasoning ontology (12) was applied to these domain instances to reason with them (13).

Once defined, this process can establish a completely automated workflow that takes a stream of XML-encoded information model instances, transforms them to OWL, and reasons with them to generate domain-level inferences.

Results

We used our methodology to generate automated workflows for two applications: (1) reasoning with radiological image annotations for tumor assessment; and (2) discovery of associations between gene mutations, drug regimens, and outcomes in HIV anti-retroviral therapy.

Reasoning with Image Annotations

The AIM Project [4] recently developed an information model that describes the semantic contents of radiological images. AIM defines an XML-encoded information model that describes anatomic structures and visual observations in the images. Information about image annotations is recorded in its information model, with the goal of enabling the consistent representation, storage, and transfer of the semantic meaning of imaging features. A variety of tools are being developed to produce image annotations in AIM format.

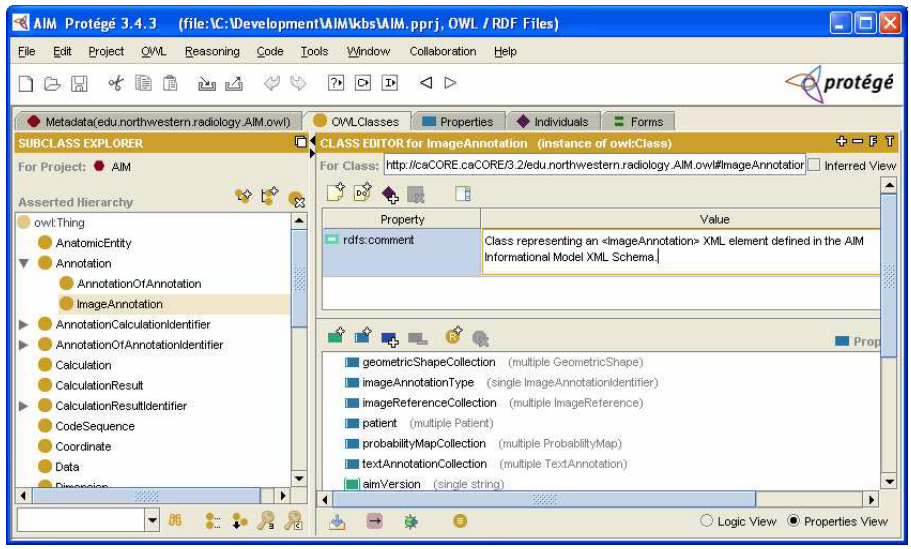

Figure 2-Screen shot of Protégé-OWL tool showing part of AIM information model encoded in OWL.

Our reasoning task employed three types of reasoning subtasks: classification, spatial reasoning, and temporal reasoning. It concentrated on classification of image findings as measurable and non-measurable using a combination of semantic information about the location and type of finding and its calculated length. These reasoning tasks were encoded using OWL and SWRL. Rules for this subtask included: classification of findings as pathologic or non-pathologic based on the imaging observation in the image annotation; temporal classification lesions as baseline or follow-up based on their temporal relationship to the start of therapy; and classification of image findings on baseline images as measurable or non-measurable based on the length of the observed mass or nodule.

We evaluated our system by defining a process to reason about cancer lesions for estimating tumor burden [11]. We used 116 AIM XML annotated images from 10 cancer patients who had serial imaging studies. Lesions in the original images were annotated in AIM format. We successfully defined a process to map image annotations encoded using AIM into OWL (Figure 2) and to reason with the resulting annotations. The image annotations were processed by our system to perform automated reasoning about the image findings.

The inferences from our system were reviewed by an oncologist, who confirmed that they were valid based on his analysis of the image annotation information encoded as instances of the AIM XML information model. In qualitative terms, the oncologist believed that our automated workflow can help streamline the process of evaluating tumor burden.

Outcome Reasoning for HIV Antiretroviral Therapy

Associations between gene mutations, drug regimens, and therapy outcomes is central in HIV therapy. In HIV research, for example, a mutation on the viral genome may be associated retrospectively to past administration of a specific drug or prospectively to the occurrence of poor clinical outcome with one or more drugs. Establishing such temporal associations may help scientists understand how certain mutations in the genome reduce drug efficacy, and can they help healthcare providers design treatment strategies.

To study drug resistance in the context of clinical care, researchers at Stanford University have developed a research system called the Stanford HIV Drug Resistance Database (HIVdb) [10]. This database contains time-stamped data on drug regimens, HIV reverse transcriptase (RT) and protease sequences, and HIV viral load collected at local clinics. Some of this information is downloadable as XML-encoded information model instances from the HIVdb website. An XML Schema for this information model has been published by the developers of HIVdb [13]. This information model describes individual patient therapies—which are termed *treatment change episodes* (TCEs)—and lists the drugs in the therapy, together with each patient's viral load response and mutation information treatment. By using information encoded in TCEs, the website can suggest ranges of suitable drug therapies.

Reasoning with TCEs requires strong temporal reasoning support. As mentioned, we have developed a temporal reasoning library for use with SWRL [12]. We used it to develop domain reasoning tasks defined in terms of these TCEs. These tasks encode several drug resistance interpretation algorithms for predicting the value of genotypic resistance test interpretation algorithms that have been described in the literature [10]. Subtasks of the reasoning task include examining patient treatment histories for particular treatment combinations, viral load patterns, and genotypic test results.

Using the HIVdb XML Schema, we defined an OWL equivalent of the information model it describes. We mapped the

information model to a domain ontology modeling TCEs and then developed a temporal reasoning module to reasoning with the resulting OWL instances. The reasoning mechanisms use OWL and SWRL, and produce high-level abstractions of patient outcomes based on a temporal analysis of their viral loads. This information can then be used for further analysis. The ultimate goal is to replicate a large subset of the therapeutic suggestion functionality of the HIVdb site. We successfully defined an automated workflow that took XML instances of the TCE information model and generated an intermediate analysis of patient outcomes.

Discussion

The increasing use of XML information models in biomedicine provides an opportunity to develop methods to automatically reasoning with the content of these models. However, XML-based information models do not typically represent information in a form that is directly suitable for reasoning—they provide a standardized interchange and storage format only. Elevating the information content from this primarily structural level to the domain level is a prerequisite to performing semantic reasoning. Using a suite of open source Semantic Web tools, we show how we have developed an approach to perform this transformation and to carry out OWL-based reasoning on information encoded in XML-based information models. We show how semantics reasoning methods were applied to these ontologies to generate domain-level inferences. Our approach establishes an automated workflow, taking XML-based information models, transforming them to an OWL domain ontology, and reasoning with the resulting information to generate inferences necessary for the domain task. We applied this workflow to perform semantic reasoning with data contained in information models in the HIV and radiological image domains.

Our approach can be used to take any XML-based information model, generate its OWL equivalent, and then reason over it to produce high-level abstractions. This approach maintains all knowledge of these transformations in OWL mapping ontologies. As a result, these mappings can be maintained at the knowledge level using standard OWL tools. Modifications to the mappings to cater for changes or extensions to the information model or domain ontologies can also be carried out using these tools. We believe that this approach provides a flexible, expandable, and robust mechanism for defining the information transformations necessary to support semantic reasoning on a large variety of biomedical data.

Acknowledgments

This research was supported in part by grant 1R01LM009607 from the National Library of Medicine.

References

[1] Shabo A, Rabinovici-Cohen S, Vortman P. Revolutionary impact of XML on biomedical information interoperability. IBM Systems Journal 45(2): 361-372, 2006.

[2] XML Schema: http://www.w3.org/XML/Schema, 2004.

[3] Dolin, RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo A, HL7 Clinical Document Architecture, Release 2, Journal of the American Medical Informatics Association 13(7):30-39, 2006.

[4] Rubin DL, Mongkolwat P, Kleper V, Supekar K, Channin DS. Medical imaging on the Semantic Web: annotation and image markup. AAAI Spring Symposium Series, Semantic Scientific Knowledge Integration, 2008.

[5] Langlotz, CP RadLex: a new method for indexing online educational materials. Radiographics, 26(6):1595-7, 2006.

[6] McGuinness D, van Hermelem F. OWL Web Ontology Language: http://www.w3.org/TR/owl-features/, 2004.

[7] SWRL: http://www.w3.org/Submission/SWRL/, 2004.

[8] Knublauch H, Fergerson RW, Noy NF, Musen MA. The Protégé OWL Plugin: An open development environment for semantic web applications. Third International Semantic Web Conference, Hiroshima, Japan, pp. 229-243, 2004.

[9] O'Connor MJ, Das A. SQWRL: a query language for OWL. OWL: Experiences and Directions (OWLED), Fifth International Workshop, Chantilly, VA, 2009.

[10]Rhee SY, Fessel WJ, Liu TF, Marlowe NM, Rowland CM, Rode RA, Vandamme AM, Van Laethem K, Brun-Vezinet F, Calvez V, Taylor J, Hurley L, Horberg M, Shafer RW. Predictive value of HIV-1 genotypic resistance test interpretation algorithms. Journal of Infectious Diseases, 200(3):453-463, 2009.

[11] Levy M, O'Connor MJ, Rubin DL. Semantic reasoning with image annotations for tumor assessment. AMIA Annual Symposium, San Francisco, CA, 2009.

[12] O'Connor MJ and Das AK. A lightweight model for representing and reasoning with temporal information in biomedical ontologies. International Conference on Health Informatics (HEALTHINF), Valencia, Spain, 2010.

[13] Stanford HIV Database Information Model Schema: http://hivdb.stanford.edu/TCEs/schema/TCE.xsd, 2009.

[14] Bohring H, Auer S. Mapping XML to OWL ontologies. Leipziger Informatik-Tage, LNI, 72:147–156, 2005.

[15]Blicer V, Leleci GB, Dogac A, Kabak Y. Providing Semantic interoperability in the healthcare domain through ontology mapping. Sigmod Record, 34(3), 2005.

Address for correspondence

Martin J. O'Connor
Stanford Center for Biomedical Informatics Research,
251 Campus Drive, MSOB X275,
Stanford, CA 94305, USA