Advances to Semantic Interoperability Through CPR Ontology Enrichment Extracting from SOAP Framework Reports

David Mendes and Irene Pimenta Rodrigues

Universidade de Évora - Portugal

Abstract

We present the work done as a contribution to use an enriched ontology as the support for Semantic Interoperability among clinicians and systems in healthcare providing environments. Clinical practice ontologies are the next generation workhorse for automatic reasoning using Semantic Web techniques and tools in the healthcare sub-domain. Ontology instance acquisition from semi-structured data that renders a full picture of the general clinical practice is crucial for solid enrichment of an Ontology that is designed to embrace the generality of information located in EHR systems. These systems communicate syntactically using HL7 standardized messaging but must evolve to semantic interoperability based in a well formed standardized semantic where CPR acts as a Knowledge Representation infrastructure. Automated acquisition is absolutely a must given the enormous amounts of information available in the mentioned sources. Recent efforts directed to solve the overwhelming complexity of HL7 V3 CDA archetype, like the greenCDA template proposal, along with computability gained with OWL DL ontologies reasoning is leading to the possibility of development of foundations for strong Clinical Decision Support tools and Computable Semantic Interoperability representations in the Semantic Web. As an intermediate step to acquisition from standardized messaging we present the ontology population/enrichment taken from the widespread framework for communication that is the SOAP (Subjective, Objective, Assessment, Plan) clinical encounters documenting system.

Keywords: Computable Semantic Interoperability; CPR Ontology; Automatic Ontology Enrichment; SOAP Framework

1 Introduction

Several technologies related to the Semantic Web as well as scientific knowledge and standardization efforts have been developing very recently at an astonishing pace that lead us to consider that we have come of age of gathering them all together and produce valuable contributions to the sub-domain of clinical practice automated reasoning. Most of the developments introduced have been maturing for years, or even decades, but finally in the last 2 years the convergence of the mentioned fields are rendering availability of usable products that will allow artificial intelligence researchers and noticeably natural language processing researchers [1] to build upon.

2 Work done

We have been developing efforts to extract clinical information from texts in Portuguese in order to represent them in computable forms able to reason about using Semantic Web tools and techniques [2]. Previously there was work done in proper selection of an adequate form Knowledge Representation (KR) suitable for the task of supporting the possibility of automated reasoning about clinical practice [3]. Meanwhile, we presented our proposal of ontology population from HL7 V2.xml messaging not yet specifying the target ontologies but
When developing our previous work \cite{3} we faced several issues of overlapping, ambiguity, non-completeness and more. Trying to figure out the availability of such an ontology suitable for our purposes, we tried to find an available standard ontology that have issues of overlapping, ambiguity, non-completeness and more. Trying to figure out the availability of such an ontology suitable for our purposes, we tried to find an available standard ontology that have isersonal nicknames for example. We have to abide to solid design foundations for proper Ontology alignment and interoperability. Well formed ontologies are able to support a variety of secondary uses not anticipated when the ontology was originally conceived \cite{6} and we may have to pick among available ontologies that have issues of overlapping, ambiguity, non-completeness and more. Trying to figure out the availability of such an ontology suitable for our purposes, we tried to find an available standard ontology according to the Ontology Realism principles enunciated in \cite{7} and with the freedom to be extendable according to anyone’s particular needs. The ontologies shall be in accordance to the OBO Foundry principles. We just try to bring together the latest Software Engineering principles to the Ontology Engineering field. With the loose coupling availability and configurable service intermixing, we picked what we could spot has low-hanging fruit to develop our systems rendering them sub-optimal but demonstrable of the validity of the concepts and easily extendable/tunable with better ontology support, finer Web Service provisioning but most of all with better clinical judgement about the “smart choices” that have to be taken to better populate the ontology given the scarce source structure of the original clinical episodes texts as seen in section 5.

4.1 Theoretical Considerations

The medical practice sub-domain we want to represent is a many faceted science that renders complexities with difficult issues to be addressed such as Temporality, Location, Granularity, High ambiguity in free text terminology, Jargon plagued with acronyms and even personal nicknames for example. We have to abide to solid design foundations for proper Ontology alignment and interoperability. Well formed ontologies are able to support a variety of secondary uses not anticipated when the ontology was originally conceived \cite{6} and we may have to pick among available ontologies that have issues of overlapping, ambiguity, non-completeness and more. Trying to figure out the availability of such an ontology suitable for our purposes, we tried to find an available standard ontology according to the Ontology Realism principles enunciated in \cite{7} and with the freedom to be extendable according to anyone’s particular needs. The ontologies shall be in accordance to the OBO Foundry principles. We just try to bring together the latest Software Engineering principles to the Ontology Engineering field. With the loose coupling availability and configurable service intermixing, we picked what we could spot has low-hanging fruit to develop our systems rendering them sub-optimal but demonstrable of the validity of the concepts and easily extendable/tunable with better ontology support, finer Web Service provisioning but most of all with better clinical judgement about the “smart choices” that have to be taken to better populate the ontology given the scarce source structure of the original clinical episodes texts as seen in section 5.

4.2 Classes to populate

We try to follow the CPR archetypes depicted in Figure 1 as close as possible when acquiring from free-text. However, some of the more complex classes like \texttt{cpr:medical_problem} are fairly hard to formulate and we have to admit that the gleaning presented in the present work renders a rather simplistic representation. The formalization of a medical practice theory however, the purpose of the CPR ontology, renders the possibility of important automated reasoning from the simple acquisition that we are proposing.

A much more systematic representation has to be used as source of information for the rendered ontologies to approach more the reality of clinical practice.
Mendes and Rodrigues | electronic Journal of Health Informatics 2012; Vol 7(x):eX

Figure 1: CPR Patient Record Archetypes

for a given clinician, medical device, service, specialty, healthcare providing institution or system. Some proposals that can contribute to this enhancement of our work are presented ahead in the Future Work section.

5 Automated acquisition from Clinical Episodes Text

As reviewed in [4] the state-of-the-art for acquisition from clinical text has enjoyed strong developments in recent years. In the mentioned paper we present a proposal for automated acquisition from HL7 messaging but here we are delving into the more generic possibility of extracting from free text present in most reporting systems used by clinicians. Going from clinical episodes text that is usually presented in a human friendly format to one adequate for computer processing involves a fair amount of text processing to handle different situations because: (1) Reports aggregate information from different clinical episodes that are not uniquely identified or not even individually dated, (2) the clinician is only identified by his/her name if any identification is made at all, (3) the information conveyed in free text is intended only to be understandable by fellow practitioners or even by the clinician him/herself making use of pragmatic jargon normally plagued with acronyms and nicknames abundant in their specific community, (4) text is profoundly intermixed with decorative elements for better legibility, normally in PDF or HTML files, (5) the clinicians natural language may be other than English without concepts defined in a foundational thesaurus like SNOMED CT or FMA for instance that don’t exist in their particular language and finally, (6) the time spanning and snapping of the processes depicted in natural language is often difficult to extract and represent formally. We intend to collect our information from SOAP reports like the de-identified sample in Figure 2.

5.1 SOAP PDF Processing

We take advantage of the fact that the report depicts a clinical encounter in a semi-structured way to manipulate into a more tractable source. The Subjective, Objective, Assessment, Plan (SOAP) framework, used to structure progress notes to facilitate problem specific, clinical decision making by physicians, is a well known, canonical structure in the medical domain. The underlying structure of the SOAP report induces some very important assumptions to be true. We find sections that can be associated with Subjective, the symptoms section S where we extract directly into a \texttt{cpr:symptom} record, medications found here are those administered only during the patient visit; Objective, the objective section O that are sign records \texttt{cpr:sign-finding} that we take as generator for \texttt{cpr:clinical_finding} in the Assessment section; Assessment, the analysis section A which are the clinical investigation acts whose outputs can be clinical artifacts to investigate things that can be consequence of any of physiological or pathological processes, and finally, Plan, the plan section P where the \texttt{cpr:therapeutic-act} can be extracted with all the timing, posology and prescriptions registered in that particular clinical encounter, medications here are prescribed for discharge [8]. Aggregating the instances collected so far we finally engage in the more complex \texttt{cpr:medical_problem} development. We now take advantage of the fact that we have to translate from jargon to English to customize our centralized TMM (Translation Memory Managers) like the Google translator toolkit\footnote{https://translate.google.com/toolkit/} or mymemory translation services\footnote{http://mymemory.translated.net/} enhanced with our own Translation Memories and Glos- saries. We use the architecture presented ahead in section “The full Software architecture picture” to do all the juggling involved in workflow processing of our source documents. We start from a PDF document, export to XML, de-identify according to determined legal ruling (like HIPAA Safe-harbor) [9], refine the clinical jargon using automated translations with the aid of CAT (Computer Aided Translation) software with previously trained Translation Memories in TMX standardized for-
Figure 2: SOAP report de-identified sample
mat and then finally convert into raw text to proceed to Semantic Annotation. The pre-processing workflow may be roughly depicted in Figure 3.

5.2 The adequate annotation workflow

A set of sequential steps must be used to go from the pure text to the extracted concept instance. Those steps workflow can be configured declaratively using the software architecture shown in the specific architecture section. The translation steps involved are: (1) Manual translation (that is indispensable for the translator tutoring) with the precise clinicians validation of their jargon adequately translated into English, (2) PDF to raw text, or to structured (XML), converting for adequate documents cleansing. The remote annotation steps are: (3) NER (Named Entity Recognition) of all the patient names, clinician names, addresses, symptoms, signs and prescriptions with all the acronyms, units, time-spanning and time-snapping involved with the usual short forms that a particular doctor usually uses. In our particular situation we maintain a local cache of the previously identified vocabulary to check exact concept matching, (4) WSD (Word Sense Disambiguation) where terms can be disambiguated without technical clinical expertise. Most of them however have to be disambiguated according to the previously acquired concepts in our controlled vocabulary, (5) EAV (Extraction of Attributes and Values) is the final pure, single language, task that has to be performed and in which we need the tooling that this paper refers to filter the concepts from the annotated text to extract concept instances, (6) Semantic annotation using the interface of BioPortal, either manually using the interactive interfaces or automatically with the Web Services available. Given the array of Web Services that can semantically annotate bio-medical concepts in English, we chose to use an evolutionary approach for use of the BioPortal annotator. We first use the annotator Web Interface after manual preprocessing for the TM tutoring and later a fully automatic workflow based in Web Services orchestration.

6 Ontology driven annotation

The annotation step is done using annotation provided by semantically aware REST annotators in BioPortal. This service renders a two step annotation process that builds upon user-selected ontologies to perform the second step of semantic annotation expansion by trying to correlate the term identified in the first step among the selected biomedical ontologies:

The Semantically Expanded Annotations are then returned by the REST Web Service in TXT, CSV or XML. With this latest option we can proceed to extract into a CPR instance. To define what elements of the returned XML are suitable for CPR instances we further enhance the XML with GRDDL entries specifying the appropriate XSLT transform to apply (2 XML lines). Unfortunately the announced web service provisioning of annotations in OWL format is not yet available from BioPortal but we have the proposed structure and that gives us the ability of developing the transforms for OWL converting as soon as that format becomes available.

7 Restricting to Clinical Practice using only SNOMED CORE

One important and pragmatic restriction that we impose is in the use of SNOMED CORE to restrain the terminology mapping. Instead of using the BioPortal SNOMED CT that would render us the choice among more than 311,000 concepts we intend to explore only a consolidated view of clinical practice terms and we now have the possibility to use only SNOMED CORE that turns our solution much more manageable while not loosing relevant terminology. Available since July 2009, the CORE (Clinical Observations Recording and Encoding) Problem List Subset of SNOMED CT was derived based on data-sets submitted by seven large scale healthcare institutions. The purpose of the CORE Project is to define a Unified Medical Language System subset that is most useful for documentation and encoding of clinical information at a summary level, such as problem list, discharge diagnosis or reasons of encounters. The most frequently used terms (covering 95% of usage volume) from these institutions are mapped to the corresponding SNOMED CT concepts where such concepts exist. The Subset contains about 5,000 SNOMED CT concepts and this downsizing permits the reasoners to classify much more effectively the resulting annotations.

This important constraint that we are imposing is a practical solution based in what views are currently available but does not take into proper account recent research that is evolving regarding issues in automated extraction like modularization and segmentation as introduced in [10].

http://bioportal.bioontology.org
Figure 3: SOAP text processing workflow

Figure 4: BioPortal Ontology Driven Annotation Workflow
7.1 Automatic transform generation

Based on the availability of CPR class archetypes and the XML structured responses from the REST Web Services, we developed XSLT mappings that are invoked through the use of GRDDL mechanism. In the case of XSLT, we just use plain XML processing but we can even use some Java based framework that has the ability to process Ontologies and GRDDL like Jena to integrate in the Java Enterprise Edition based system presented in the next section.

7.2 The full software architecture picture

As presented in [11], we build our proposal based on a lightweight messaging bus that we call Clinical Practice – Enterprise Service Bus (CP-ESB). This RESTful-based hub is responsible for orchestrating all the communications between Web Service invocations allowing a high degree of customization and plug-and-play ability that renders our proposal very flexible and future proof:

Reviewing the above referred points from SOAP we can now summarize in the following table the complete acquisition points that fit in the timeline for clinical process acquisition proposed by Schueermann et al. in 2009 [12]. Each identified point generates a gleaning entry in soap_to_cpr.xslt transform file summarized in Table 1.

As explained in the “Classes to populate” section we can only collect a thin representation when using our SOAP approach but the application of this methodology by the enormous number of clinicians that use this globally widespread framework delivers interesting results like the mere inferencing of the actual relations between collected instances from S.O.A and the consequences described in P.

8 Path to facilitate CSI

CSI is not dependent of any particular implementation, data structure or archetype in any of the different points that have to share the knowledge. What is important is a shareable meaning foundation that has to be agreed by all. We have previously argued [14] that an important contribution is brought by BRIDG. The Biomedical Research Integrated Domain Group (BRIDG) Model is a collaborative effort engaging stakeholders from the Clinical Data Interchange Standards Consortium (CDISC), the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), the National Cancer Institute (NCI) and its Cancer Biomedical Informatics Grid (caBIG®), and the US Food and Drug Administration (FDA). The BRIDG model is an instance of a Domain Analysis Model (DAM). The goal of the BRIDG Model is to produce a shared view of the dynamic and static semantics for the domain of protocol-driven research and its associated regulatory artefacts. In the mentioned paper we sustain that in Healthcare, every system involved must agree in the meaning of a clinical concept. This is the fundamental concept of Shared Semantics that has to be realized to obtain CSI in the Health sub-domain of knowledge. We present the HL7 RIM inadequacy for ontology mapping and how to circumvent it using the BRIDG DAM. This approach will serve as guideline to alleviate the evident differences between CPR structure and the underlying models structure like the RIM or SOAP. Several options that are taken pursue the DAM.

9 Future work

In 2007 the work done under the W3C - Semantic Web Healthcare and Life Sciences Interest Group auspices rendered amazing contributions done with the most interesting technological landscape available at the time. Some realities that have since evolved and we could use for the present work are:

(a) The Problem-Oriented Medical Record Ontology that was started to be developed in 2006 was the target of the fruitful GRDDL efforts that appeared during 2007 in which base we are now working. This efforts however were rdfs based for the OWL developments were still incipient. Only in

http://www.bridgmodel.org/
http://www.w3.org/2001/sw/hcls/
November 2009 W3C standardized a OWL based ontology called the CPR that we use in our work.

(b) For a given HL7 Clinical Document Architecture (CDA) document the GRDDL gleaning presented in 2007 with it associate transform[9] creates one cpr v0.5 instance with one cpr:patient.record with as many cpr:screening-acts and associated cpr:clinical-descriptions as episodes are referred in the CDA source. This approach, though impressive technology, lacks the granularity that can be rendered now with further developments that CPR has gained in its latest versions.

Taking advantage of the developments done so far we intend to build upon what was then achieved and propose work to be developed like:

9.1 Extend our proposal to use GRDDL for gleaning HL7 documents into CPR

The same architecture that we are proposing for gleaning from BioPortal’s Open Biomedical Annotator (OBA) output may be, as well, used to extract resources from CDA documents compliant with RIM V3. However it seems obvious that the complexity that the specification conforms could easily “hinder the trees and not show the forest”. Many recent proposals are being developed to apply restricted templates to make RIM tractable. The most notorious, that we intend to make reference to here, is the greenCDA[10] project that already has delivered results. In fact, the greenCDA specification has been officially balloted via HL7 and the version 1.0 is already available for download[11].

9.2 Using only refined templates of HL7 V3 like greenCDA instead of the full CDA

The idea was that the CDA was capable of representing any aspect of the medical record for any purpose. That sounds like a noble idea but in practice it creates a fixed overhead for even the simplest data exchange. The greenCDA project has developed a pragmatic methodology for creating simplified schemas that can be transformed directly to or from normative CDA. The initiative can be briefly described as a simple to use, XML construct that incorporates structured and unstructured clinical summary information. Simpler than CDA but full featured. There are tools available to convert CDA into Green CDA and also additional tools are being developed to enable easy creation of Green CDA constructs by navigating the RIM, selecting attributes, and selecting associations to consolidate to make the XML flatter. We feel that the development of XSLT for GRDDL processing from greenCDA documents is a very viable process of rendering a quasi complete picture of clinical practice extracting from HL7 messaging. We already have available the XML-Schema and transforms for back and forth CDA conversion and very easily will develop the OBA annotator output of greenCDA documents to CPR GRDDL transforms.

9.3 Restrict CPR to be OWL DL well formed or even further

Unfortunately, CPR is not currently OWL DL syntax conformant. For the current reasoners to work effectively with OWL based ontologies they have to be OWL DL to be validatable in linear time. Although our clinical practice ontologies are restricted in their size most of the times, the reasoners check the validity a-priori and tend to refuse to work when some non-conformities are found. One interesting line of research may be the adequate restriction of CPR to be OWL DL conformant without lost of the needed expressiveness for clinical practice representation. We found even arguable that

Table 1: XSLT transforms for XML OBA Annotation to CPR instance

<table>
<thead>
<tr>
<th>Instance Type</th>
<th>SOAP Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr:person</td>
<td>H: Header (if not de-identified)</td>
</tr>
<tr>
<td>cpr:patient</td>
<td>H</td>
</tr>
<tr>
<td>cpr:symptom</td>
<td>S</td>
</tr>
<tr>
<td>cpr:sign-finding</td>
<td>O</td>
</tr>
<tr>
<td>cpr:clinical-finding</td>
<td>A</td>
</tr>
<tr>
<td>cpr:clinical_diagnosis</td>
<td>A</td>
</tr>
<tr>
<td>cpr:therapeutic-act</td>
<td>P</td>
</tr>
<tr>
<td>cpr:medical_problem</td>
<td>All</td>
</tr>
</tbody>
</table>

9http://www.w3.org/2001/sw/grddl-wg/td/hl7-rim-to-pomr.xslt
with adequate restrictions, like elimination of transitivity [13] a Horn SHIQ ontology can be developed from CPR and then we reach the possibility of applying Consequence-Driven Reasoners that have been shown recently to be very attractive for dealing with huge datasets.

10 Conclusion

We presented our proposal for using the enriched CPR ontology, automating the acquisition with resource to SOAP framework reports, as support for Computable Semantic Interoperability. Being a framework directed to human interpretation it lacks the completeness that can be achieved with more formal ways of representing clinical encounters directed to computer consumption like the various standardized protocols and archetypes developed for computer to computer interaction through messaging like ISO HL7 27931:2009 or the Archetypes in HL7 V3 RIM. We demonstrated the workflow to evolve from a raw SOAP report in natural language to a CPR instance creation both manually and automatically through the use of the available Web Services that “ontology driven” annotate with resource of “feeder ontologies” that we criteriously select. Our final architecture has the possibility to automatically process the source documents into the ontology through the suggested software components that is easily expanded/refined to incorpore new sources of clinical information, enhanced ways of annotation and different target ontologies in a plug-and-play manner.

References


Conflicts of Interest

None declared.

Correspondence

David Mendes
+351 963564161
dmendes@uevora.pt