

Semantic Representation of Evidence-Based Clinical Guidelines

Zhisheng Huang¹(✉), Annette ten Teije¹(✉), Frank van Harmelen¹,
and Salah Ait-Mokhtar²

¹ Department of Computer Science, VU University Amsterdam, Amsterdam,
The Netherlands

{huang, annette, Frank.van.Harmelen}@cs.vu.nl

² Xerox Research Centre Europe, Meylan, France
Salah.Ait-Mokhtar@xrce.xerox.com

Abstract. Evidence-based Clinical Guidelines (EbCGs) are document or recommendation which have been created using the best clinical research findings of the highest value to aid in the delivery of optimum clinical care to patients. In this paper, we propose a lightweight formalism of evidence-based clinical guidelines by introducing the Semantic Web Technology for it. With the help of the tools which have been developed in the Semantic Web and Natural Language Processing (NLP), the generation of the formulations of evidence-based clinical guidelines become much easy. We will discuss several usecases of the semantic representation of EbCGs, and argue that it is potentially useful for the applications of the semantic web technology on the medical domain.

1 Introduction

Clinical guidelines (CGs) are recommendations on the appropriate treatment and care of people with specific diseases and conditions. Evidence-based Clinical Guidelines are that the document or recommendation has been created using an unbiased and transparent process of systematically reviewing, appraising, and using the best clinical research findings of the highest value to aid in the delivery of optimum clinical care to patients. Clinical guidelines have been proved to be valuable for clinicians, nurses, and other healthcare professionals in their work. Of course, clinical guidelines would not replace the medical knowledge and skills of healthcare professionals.

Computerized Clinical Guidelines, alternatively called Computer-Interpretable Guidelines (CIGs), implement the guidelines in computer-based decision support systems. Computerized clinical guidelines are expected to improve the acceptance and application of guidelines in daily practice for healthcare professionals, because their actions and observations can be monitored by a semi-automatic way with the support of a computer-based decision support system. When the decision support system has detected that a guideline is not followed, an advice would be generated by the system [4]. Various computerized clinical guidelines as well as decision support systems that incorporate these guidelines

have been developed. Those main standards of computerized clinical guidelines or medical decision support languages are: the Arden Syntax¹, PROforma [6, 7], Asbru [14], EON [16], GLIF [12, 13]².

In this paper, a main concern is the semantic interoperability, which is usually achieved by mapping concepts in the specification of guidelines to standard terminologies and domain ontologies. Both EON and GLIF emphasize its integration with medical terminologies and domain ontologies. Neither PROforma and nor Asbru is developed for the integration of medical terminologies or domain ontologies.

In this paper, we propose a semantic representation of evidence-based clinical guidelines as a lightweight formalism of EbCGs. Those evidence-based guidelines are represented by using the RDF/RDFS/OWL standards, which have been widely used in the Semantic Web Technology. We have used XMedlan, an NLP tool [1], to generate the semantic data of clinical guidelines. The proposed formalism is designed to serve for the task of the guideline update in the European 7th framework project EURECA [3]³.

Compared with existing other formalisms of Computer-Interpretable Guidelines, which usually require a lot of manual processing for the generation of the formulation, this lightweight formalism of evidence-based clinical guidelines has the advantage that they can be generated quite easy by using the tools that have been developed in the NLP and Semantic Web. We will argue that this semantic representation of evidence-based clinical guidelines have some novel features, and can be used for various application scenarios in the medical domain.

The main contribution of this paper is:

1. propose a framework of semantic representation of evidence-based clinical guidelines,
2. show how to use an NLP tool to generate semantic data of clinical guidelines,
3. present several use cases of the lightweight formalism of evidence-based clinical guidelines for the semantic operability.

This paper is organized as follows: Sect. 2 introduces the general ideas of evidence-based clinical guidelines. In Sect. 3 we propose a semantic representation of evidence-based clinical guidelines by using the semantic web technology. Section 4 describes the NLP tools that have been used to convert the textual clinical guidelines into semantic data. Section 5 discusses several use cases of the semantic representation of evidence-based clinical guidelines. Section 6 discusses the related work, future work, and make the conclusions.

2 Evidence Based Clinical Guidelines

As we have discussed above, Evidence-based Clinical Guidelines are a series of recommendations on clinical care, supported by the best available evidence in

¹ <http://www.hl7.org/special/Committees/arden/index.cfm>

² http://web.squ.edu.om/med-Lib/med/net/e-pathways-net/Docs/GLIF3.TECH_SPEC.pdf

³ <http://eurecaproject.eu>

the clinical literature. In evidence-based clinical guidelines, the answers to the fundamental questions are based on published scientific research. The articles selected were evaluated by an expert in methodology for their research quality, and graded in proportion to evidence using the classification system described in [11]. In [11], the following classification of research results are proposed on level of evidence.

- A1. Research on the effects of diagnostics on clinical outcomes in a prospectively monitored, well-defined patient group, with a predefined policy based on the test outcomes to be investigated, or decision analysis research into the effects of diagnostics on clinical outcomes based on results of a study of A2-level and sufficient consideration is given to the interdependency of diagnostic tests.
- A2. Research relative to a reference test, where criteria for the test to be investigated and for a reference test are predefined, with a good description of the test and the clinical population to be investigated; this must involve a large enough series of consecutive patients; predefined upper limits must be used, and the results of the test and the “gold standard” must be assessed independently. Interdependence is normally a feature of situations involving multiple diagnostic tests, and their analysis must be adjusted accordingly, for example using logistic regression.
- B. Comparison with a reference test, description of the test and population researched, but without the other features mentioned in level A.
- C. Non-comparative trials
- D. Opinions of experts, such as guideline development group members.

Furthermore, the conclusions in the guidelines, alternatively called *guideline items*, are annotated with an evidence level. Based on the medical literature, one or more relevant conclusions are made for each section. The most important literature is listed according to the level of evidential strength, allowing conclusions to be drawn based on the level of evidence. All the medical literature included in the conclusion is described in the bibliography. The classification of conclusions are based on literature analysis. The following evidence levels are proposed in [11].

- Level 1. Based on 1 systematic review (A1) or at least 2 independent A2 reviews.
- Level 2. Based on at least 2 independent B reviews
- Level 3. Based on 1 level A2 of B research, or any level C research
- Level 4. Opinions of experts, such as guideline development group members

Here are some examples of evidence-based clinical guidelines:

Level 1

Breast cancers detected through regular breast self-examination have no better prognosis than breast cancers detected by other means.

A1 Ksters 2003, Elmore 2005, Weiss 2003, Nelson 2009.

Level 1

A self reported lump by the woman is positively associated with an actual mass being present.

A2 Barlow 2002, Lumachi 2002, Aiello 2004.

Each guideline consists of an evidence level, a guideline statement, and their references with the classification of the research results (e.g. A1 or A1).

3 A Lightweight Formalism of Evidence-Based Clinical Guidelines

The advantage of Computer-Interpretable Guidelines is that they implement the guidelines in computer-based decision support systems. However, the generation of the formulations of the guidelines according to the existing formalisms requires a lot of manual processing. That leads to different formalisms of Computer-Interpretable Guidelines, which can be ranged from a high-level representation (i.e., more expressive and more logic-oriented), such as Asbru to a low-level representation such as the Arden Syntax. There may exist different requirements on evidence-based clinical guidelines from different perspectives or application scenarios. In this paper, we are concerned with the following requirements on evidence-based clinical guidelines:

- **Structured Data:** Existing clinical guidelines are usually available at the textual format (i.e., a pdf file or word document). They are not computer-Interpretable. We have to convert those textual data into the structured data. Structured data has the advantage of being easily entered, stored, queried and analyzed.
- **Semantic Interoperability.** Semantic interoperability is the ability of computer systems to exchange data with unambiguous, shared meaning. Semantic interoperability is therefore concerned not just with the packaging of data (syntax), but the simultaneous transmission of the meaning with the data (semantics). This is accomplished by adding data about the metadata, linking each data element to a controlled, shared vocabulary.⁴
- **Reasoning Support:** The formalism should be enriched with the knowledge technology, in particular, support for reasoning over the knowledge which have been contained in clinical guidelines.
- **Generation Convenience:** The generation of the formulation of clinical guidelines from the existing textual documents are usually time-consuming, because it requires a lot of manual processing. Although a natural language processing (NLP) tool is helpful to improve the efficiency of the generation, converting those information which have been obtained by using the NLP tool into a high-level formalism is still needed to be done by professional people.
- **Evidence-oriented Representation:** The formalism of evidence-based clinical guidelines should be convenient to represent different fine-grained levels of evidences.

⁴ http://en.wikipedia.org/wiki/Semantic_interoperability

Considering the requirement above, we made the following design decision. There are different data models to structure a data. They can be just simply structured as a relational database, or to be an XML file. Consider the requirements on the semantic interoperability and reasoning support, we prefer using the RDF/RDFS/OWL data format, because they have been widely used as a solution to the semantic interoperability and reasoning. Although the existing OWL reasoning may not be powerful enough to cover all the aspects of clinical guidelines such as reasoning about actions, description of uncertainty, temporal and spatial processing, workflow processing, and others, we should make a trade-off between the expressibility of high-level representation and the convenience of generation. Thus, in this paper, we propose a RDF/OWL-based formalism for the semantic representation of evidence-based clinical guidelines. For the requirement on evidence-orientation, we design RDF-based terminologies (thus a lightweight ontology) to express clinical evidences, so that those concepts can be used to represent various evidence information in clinical guidelines.

The semantic representation of evidence-based clinical guidelines consists of the following sections:

- Heading. The heading section of the guidelines provide the basic description of the information such as the title, published time, version number, and provenance.
- Body. The body section of the guidelines provide the main description of guidelines and their evidences. The body section consists of a list of guideline items (i.e., an evidence-based guideline statement), which contains the evidence information and the RDF/OWL-representations of a single guideline statement.
 - Evidence description. It provides the formal description of the evidence (i.e., evidence level and its references which use the Dublin core format, the standard metadata to represent a publication.
 - Guideline description. It provides the RDF/OWL description of the guideline statement.

We have used the XMedlan NLP tool to generate the semantic statements of guideline description. In the next section, we will discuss the features of XMedlan and how to use this NLP tool to generate the RDF/OWL statements for clinical guidelines.

Here are the examples of the semantic representation of evidence-based clinical guidelines in the RDF Triples:

The heading:

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dc: <http://purl.org/dc/1.1#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sct: <http://wasp.cs.vu.nl/sct/sct#>
PREFIX sctid: <http://wasp.cs.vu.nl/sct/id#>
```

```

PREFIX ctec: <http://eurecaproject.eu/ctec/>
sctid:g1002-zsh140412 rdf:type sct:EvidenceBasedGuidelines.
sctid:g1002-zsh140412 dc:title "Dutch Breast Cancer Guideline".
sctid:g1002-zsh140412 dc:creator "NABON".
sctid:g1002-zsh140412 sct:publicationYear "2012".

```

For each guideline item, we have the following statements of the evidence description:

```

sctid:g1002-zsh140412 sct:hasConclusions sctid:g1002-zsh140412_1 .
sctid:g1002-zsh140412_1 rdf:type sct:GuidelineConclusions.
sctid:g1002-zsh140412_1 sct:about
    "Regular breast self-examination as a screening method".
sctid:g1002-zsh140412_1 sct:hasGuidelineItem sctid:g1002-zsh140412_1_1 .
sctid:g1002-zsh140412_1_1 sct:hasEvidenceLevel "1"^^xsd:integer .
sctid:g1002-zsh140412_1_1 sct:hasReferences sctid:g1002-zsh140412_1_1ref .
sctid:g1002-zsh140412_1_1ref sct:hasReference sctid:g1002-zsh140412_1_1ref1 .
sctid:g1002-zsh140412_1_1ref1 sct:hasReference sctid:g1002-zshref-Ksters2003 .
sctid:g1002-zsh140412_1_1ref1 sct:evidenceClassification "A2".
sctid:g1002-zsh140412_1_1ref sct:hasReference sctid:g1002-zsh140412_1_1ref2 .
sctid:g1002-zsh140412_1_1ref2 sct:hasReference sctid:g1002-zshref-Elmore2005 .
sctid:g1002-zsh140412_1_1ref2 sct:evidenceClassification "A2".
sctid:g1002-zsh140412_1_1ref sct:hasReference sctid:g1002-zsh140412_1_1ref3 .
sctid:g1002-zsh140412_1_1ref3 sct:hasReference sctid:g1002-zshref-Weiss2003 .
sctid:g1002-zsh140412_1_1ref3 sct:evidenceClassification "A2".
sctid:g1002-zsh140412_1_1ref sct:hasReference sctid:g1002-zsh140412_1_1ref4 .
sctid:g1002-zsh140412_1_1ref4 sct:hasReference sctid:g1002-zshref-Nelson2009 .
sctid:g1002-zsh140412_1_1ref4 sct:evidenceClassification "A2" .
sctid:g1002-zsh140412_1_1 sct:hasRelations _:e5 .

```

which describes the guideline items with their evidence levels and references.

The following RDF N-Triples relations are extracted from the guideline statement using the NLP tool:

```

_:e1 <http://eurecaproject.eu/ctec/isA> <http://eurecaproject.eu/ctec/diagnosis> .
_:e1 <http://eurecaproject.eu/ctec/hasObject> _:e2 .
_:e2 <http://eurecaproject.eu/ctec/hasTerm> "cancers".
_:e2 <http://eurecaproject.eu/ctec/hasCUI> "C0006826".
_:e2 <http://eurecaproject.eu/ctec/isA>
    <http://eurecaproject.eu/ctec/disease_or_syndrome> .
_:e3 <http://eurecaproject.eu/ctec/isA>
    <http://eurecaproject.eu/ctec/diagnosis> .
_:e3 <http://eurecaproject.eu/ctec/hasObject> _:e4 .
_:e4 <http://eurecaproject.eu/ctec/hasTerm> "cancers".
_:e4 <http://eurecaproject.eu/ctec/hasCUI> "C0006826".
_:e4 <http://eurecaproject.eu/ctec/isA>
    <http://eurecaproject.eu/ctec/disease_or_syndrome> .
_:e5 <http://eurecaproject.eu/ctec/hasFragment> _:e1 .
_:e5 <http://eurecaproject.eu/ctec/hasFragment> _:e3 .
_:e6 <http://eurecaproject.eu/ctec/hasTerm> "Breast".
_:e6 <http://eurecaproject.eu/ctec/hasCUI> "C0006141|C1268990".
_:e5 <http://eurecaproject.eu/ctec/hasFragment> _:e6 .
_:e5 <http://eurecaproject.eu/ctec/isA> <http://eurecaproject.eu/ctec/EC> .

```

```
_:e5 <http://eurecaproject.eu/ctec/hasText> "Breast cancers detected through
regular breast self-examination have no better prognosis than breast cancers
detected by other means." .
```

which states the guideline statement (i.e., the text of the guideline) and the relation extractions from the statement. They provide the detailed RDF description of the guideline statement and their annotation with the concepts in UMLS, a well-known meta-thesaurus of medical terms developed by the National Library of Medicine [10] and offering mappings to most of widely used medical terminologies such as SNOMED-CT, NCI, LOINC, etc.

4 NLP Tool

XMedlan is the Xerox linguistic-based module of the relation extraction system of the EURECA project [1]. The main characteristic of this component is that it uses a linguistic parser [2] to perform rich linguistic analysis of the input text. Figure 1 depicts the processing steps of the NLP extraction: after an optional structure analysis of the input text document, the linguistic parser annotates each sentence with rich linguistic features and structures, which then serve as the basis of the extraction of relations and attributes in the form of triples.

4.1 Linguistic Analysis

First, the text is tokenized into a sequence of tokens, each token is looked up in a lexicon and assigned all its possible morpho-syntactic categories and features, and possibly additional semantic features. Ambiguous tokens (with more than one possible syntactic category) are disambiguated with a part-of-speech (POS) tagger leveraging the left and right contexts of the tokens. A medical concept identifier recognizes mentions of medical concepts in the sequence of tokens, and annotates them with their UMLS unique concept identifiers (CUIs) and semantic types (Anatomical_Structure, Clinical_Drug, Disease_or_Syndrome, etc.). Other types of mentions, not specific to the medical domain, are also recognized: time expressions and measures. During the syntactic parsing phase, sub-sequences of tokens and concept mentions are grouped into syntactic constituents, called chunks, by the parser: Noun Phrases, Verb Phrases, etc. Most importantly, tokens and concepts mentions are linked with syntactic dependency relations, e.g. subject, direct object, noun modifier, etc. Finally, the linguistic parser performs a local and partial semantic analysis of the input sentence and produces semantic annotations. Examples of such semantic annotations include negation on concepts, and the structural representation of measures with the following relations: *hasValue*, *hasMin*, *hasMax* and *hasUnit*.

4.2 Extraction of Relations and Attributes

All the annotations produced by the linguistic analyzer are exploited by the relation extraction engine to identify relations and attributes of concepts and

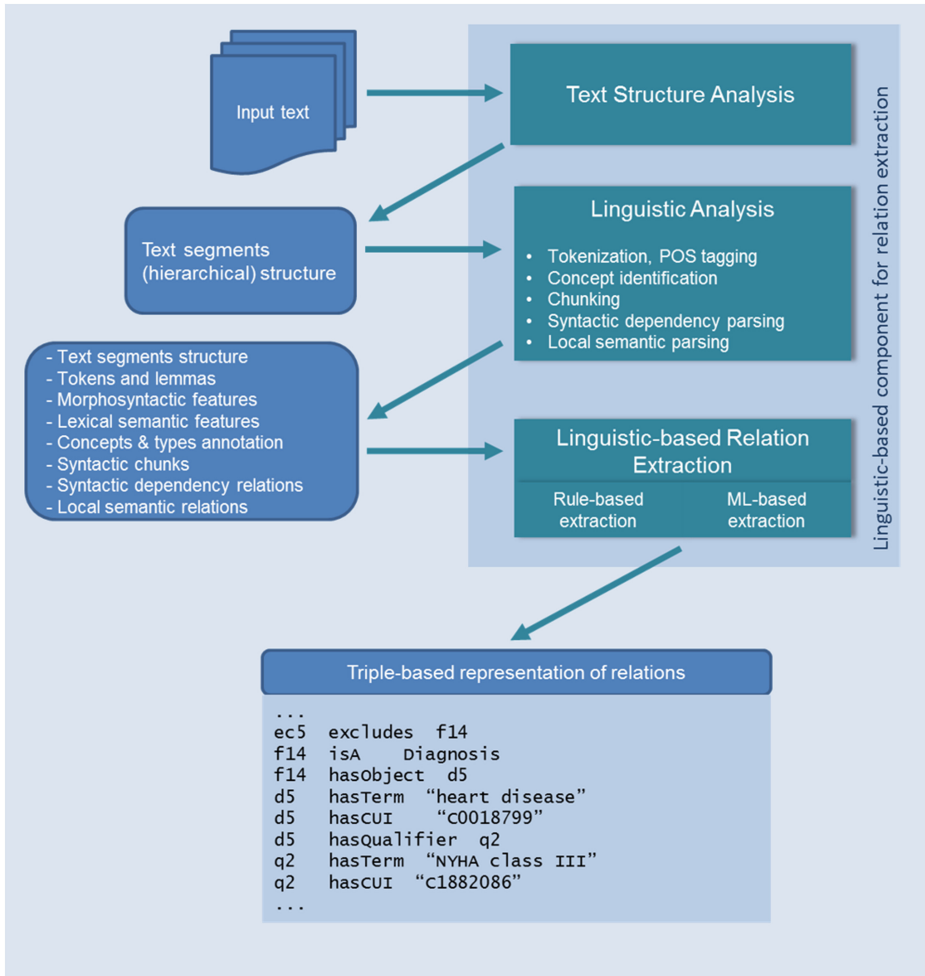


Fig. 1. Block diagram of the architecture of the Xerox linguistic-based relation extraction component

entities in the input text. These relations and attributes are expressed as triples, i.e. typed binary relations in the form $\langle Subject, Property, Object \rangle$, and can be serialized in the RDF N-Triples format.

The extraction engine has a rule-based and a machine learning sub-component. Currently we use only the former, as the implementation of the latter is not yet finalized. The rule-based sub-component is a rule interpreter implemented on top of a Prolog engine. It takes an (ordered) set of extraction rules and applies each of them to each sentence of the input text. The rules are not exclusive, several of them can succeed on the same input sentence and yield additional, and possibly inconsistent, triples: in its current state, the tool does not check the semantic

consistency of extracted triples and fully relies on the quality of the rules developed by the user. In the following paragraphs, we describe in more details extraction rules and the way they apply on input sentence to produce triples.

Extraction Rules. An extraction rule is a heuristic that has *conditions* on the linguistic features and structures produced by the linguistic analyzer (see previous sub-section), and *actions*, which consist in the creation of triples that are added to the context if the conditions are satisfied. Note that UMLS-related triples (hasTerm, hasCUI) are automatically added by the rule engine whenever a node that is a UMLS concept is referred to for the first time in a newly created triple. The scope of the rule conditions is global, i.e. a rule can express conditions on linguistic annotations of any sentence or text segment in the input document. Furthermore, the conditions can also check the existence of triples created by the previous successful execution of rules on the current sentence, or on previous sentences within the same document.

Example of an Extraction Rule. Let's assume we want to extract instances of a class *Laboratory_or_Test_Result*, which has the following properties: *hasObject*, the value of which is the thing being measured in the laboratory result, *hasValue* or *hasMaxValue* or *hasMinValue*, which give the quantitative result, and *hasProcedure*, the value of which is the laboratory procedure used to obtain the test result. Hence, the following input text “*Ventricular ectopics less than 4/min on EKG*” is represented with the following set of triples:

```
_:e1 <http://eurecaproject.eu/ctec/hasUnit> "/min".
_:e1 <http://eurecaproject.eu/ctec/hasQuant> "4".
_:e2 <http://eurecaproject.eu/ctec/hasMaxValue> _:e1 .
_:e2 <http://eurecaproject.eu/ctec/hasObject> _:e3 .
_:e3 <http://eurecaproject.eu/ctec/hasTerm> "Ventricular ectopics".
_:e3 <http://eurecaproject.eu/ctec/hasCUI> "C0151636|C0488470".
_:e2 <http://eurecaproject.eu/ctec/isA> <http://eurecaproject.eu/ctec/
      laboratory_or_test_result> .
_:e2 <http://eurecaproject.eu/ctec/hasProcedure> _:e4 .
_:e4 <http://eurecaproject.eu/ctec/hasTerm> "EKG".
_:e4 <http://eurecaproject.eu/ctec/hasCUI> "C1623258" .
```

and the (simplified) rule example below would trigger the creation of 3 triples representing an instance of *Laboratory_or_Test_Result*, *fact1*, with the *hasObject* and *hasProcedure* properties populated:

RULE Conditions

- Node1, a concept of type *Clinical_Attribute*
- Node2, of type *Quantitative_Value*
- Node3, a concept of type *Laboratory_Procedure* or *Diagnostic_Procedure*
- A syntactic dependency relation between Node1 and Node2
- A syntactic dependency relation between Node1 and Node3

RULE Actions:

```
fact1 isA Laboratory_or_Test_Result
fact1 hasObject Node1
fact1 hasProcedure Node3
```

Application of Extraction Rules. An extraction rule applies successfully on an input sentence if all its conditions are satisfied. When a condition is evaluated, its free variables are instantiated with nodes from the input sentence and/or current set of triples. For instance, in the rule example above, the first condition is satisfied as many times as there are medical terms of type *Clinical_Attribute* in the input sentence, and variable *Node1* is successively assigned values that corresponds to those terms. For each possible value of *Node1*, the second condition is evaluated and similarly, variable *Node2* will successively be assigned to the terms of type *Quantitative_Value*, if any. When the fourth condition is considered (existence of a syntactic dependency relation between *Node1* and *Node2*), variables *Node1* and *Node2* are not free and the condition evaluates to true at most once, including if there are multiple dependency relations between the two nodes. In other words, the whole conditions of a rule are satisfied (and its actions are run) **for each distinct set** of variable instantiations that make every individual condition evaluate to true. And for each distinct set of variable instantiations satisfying the conditions, the rule actions are run, i.e. each triple in the rule actions is created.

Rule Order. A rule condition can consist in requiring the existence of a triple produced by a previous rule that applied successfully on current input. This means rules apply sequentially and their order is important: extraction rules for simple semantic classes come first. Simple semantic classes are classes with literal property values or property value constraints that strongly discriminate them within the conceptual schema, which allows for the expression of discriminative rule conditions, leading to unambiguous or weakly ambiguous instantiations from input sentences. An example of a simple class is *Quantitative_Value*, having *hasQuantity* and *hasUnit* properties with literal values: whenever an input text contains a numerical value and a measurement unit linked with a syntactic dependency, these two “concept” occurrences can unambiguously instantiate the object of the aforementioned properties, yielding an instance node of class *Quantitative_Value*, with two triples. The instance node can then be used in the next extraction rules as the object of a property of a more complex semantic class (e.g. *Laboratory_or_Test_Result*, in the rule example above).

Development of Extraction Rules. The starting point for creating rules is the definition of the semantic representation of the information we want to extract in a use case, i.e. the ontology or conceptual schema that we want to “populate” from text content: a set of classes and their properties. The properties are relation or attribute types that link instances of the classes. For example, for the extraction of eligibility criteria in clinical trials [1], we defined classes

such as *Diagnosis*, *Laboratory_or_test_result*, *Age*, *Gender*, *Treatment*, etc. The *Diagnosis* class has two main properties: *hasObject*, the value of which is the disease being diagnosed, and *hasProcedure*, the value of which is the diagnostic procedure employed in the diagnosis. Each extraction rule is dedicated to the extraction of one or more relation types defined in the conceptual schema of the use case for which the rule set is created. Hence, the rules are highly dependent on the domain (medical), and even dependent on the use case: the proportion of rules that can be reused for another use case is determined by the proportion of classes and properties that are common to the conceptual schemas of the two use cases: as an example, class *Diagnosis* can be relevant both in clinical trial use cases and for the structuring of clinical guidelines.

In the current version, we are using the initial set of rules developed for the extraction of clinical trial eligibility criteria (CTEC) [1]. We plan to enrich this initial rule set and adapt it to the extraction of relations specific to the semantic representation of evidence-based clinical guidelines.

5 Implementation and Feasibility

5.1 Implementation

We have implemented a tool to generate the semantic representation of evidence-based clinical guidelines. We select the Dutch breast cancer guidelines version 2.0, which has been published in 2012, as test data. The transformation consists of the following processing:

- XML document generation. We create an XML document which contains the conclusions of the guidelines which have been marked with the evidence. Since the existing draft of the guidelines are in the pdf textual format, we have developed a tool which can extract the conclusions from a clinical guideline if those conclusions are stated with a textual pattern shown in the example above. Of course, we can also generate the XML document manually by copying and pasting the corresponding documents.
- Evidence statement generation. We use the XSLT tool to convert the XML document into a set of RDF statements which corresponding with the evidence description.
- Guideline statement generation. We use the NLP tool to generate the RDF statements for each guideline statement.

We are implementing a component of evidence-based clinical guidelines in the SemanticCT system, a semantics-enable system for clinical trials⁵ [8]. The goals of SemanticCT are not only to achieve interoperability by semantic integration of heterogeneous data in clinical trials, but also to facilitate automatic reasoning and data processing services for decision support systems in various settings of clinical trials. SemanticCT is built on the top of the LarKC (Large Knowledge Collider) platform⁶, a platform for scalable semantic data processing [5, 17].

⁵ <http://wasp.cs.vu.nl/sct/>

⁶ <http://www.larkc.eu>

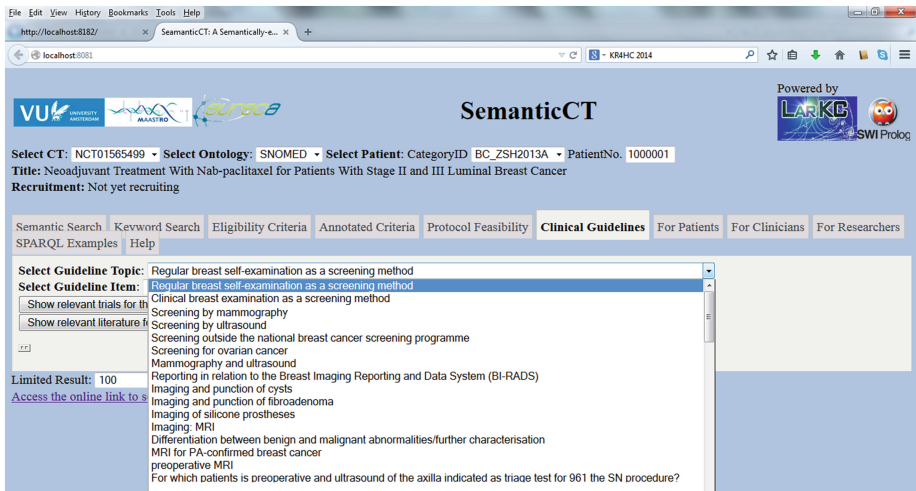


Fig. 2. The guideline component in SemanticCT

The SemanticCT management component manages the SPARQL endpoint which is built as a SemanticCT workflow which consists of a generic data processing and reasoning plugin in the LarKC platform. That generic data processing and reasoning plug-in provides the basic reasoning service over large scale semantic data, like RDF/RDFS/OWL data.

SemanticCT provides the interface of semantic search, so that a user can post SPARQL queries to obtain the results. For the users who have no any background knowledge of the Semantic Web, they can use the graphical interface to use the system for the services. A screenshot of the interface of the guideline component in SemanticCT is shown in Fig. 2.

5.2 Feasibility

Different from the existing expressive formalisms of Computer-Interpretable Guidelines such as Asbru, the lightweight formalism of the evidence-based clinical guidelines may not be efficient to be used directly for the workflow processing in clinical decision making systems. In this paper, our main concern is the semantic interoperability.

The guideline statements of the semantic representation of guidelines have been annotated with the well-known medical terminologies/ontologies such as UMLS. Although the mappings among various terminologies are usually not easy, the annotations with the same ontology like that has been done in the XMedlan provides the way to connect different data resources with the same concept annotation. It also provides the possibility for querying a SPARQL endpoint on a triple store.

Below we give a number of example queries based on our lightweight formalisation of evidence-based clinical guidelines. The first example is a SPARQL

query to list a set of concepts which have been used in the annotation of the guideline statement.

```
PREFIX sct: <http://wasp.cs.vu.nl/sct/sct#>
PREFIX ctec: <http://eurecaproject.eu/ctec/>
select distinct ?id ?text ?conceptid ?term
where {
?s ctec:hasText ?text.
?s1 sct:hasRelations ?s.
?s1 sct:hasGuidelineItemID ?id.
?s ctec:hasFragment ?e1.
?e1 ctec:hasTerm ?term.
?e1 ctec:hasCUI ?conceptid.}
ORDER BY ?id
```

Another example of the SPARQL query is to list two guideline statements which have been annotated with the same concept:

```
PREFIX ...
select distinct ?id1 ?text1 ?id2 ?text2
where {
?s ctec:hasText ?text1.
?s1 sct:hasRelations ?s.
?s1 sct:hasGuidelineItemID ?id1.
?s ctec:hasFragment ?e1.
?e1 ctec:hasTerm ?term.
?e1 ctec:hasCUI ?conceptid.
?e2 ctec:hasCUI ?conceptid.
?e2 ctec:hasTerm ?term.
?s2 ctec:hasText ?text2.
?s2 ctec:hasFragment ?e2.
?s3 sct:hasRelations ?s2.
?s3 sct:hasGuidelineItemID ?id2.
FILTER (!(?e1=?e2)).}
ORDER BY ?id
```

We are also interested in the semantic operability from different data sources. For example, the following query can be used to check the connection between evidence-based guidelines and clinical trials.

```
PREFIX ...
select distinct ?guidelineid ?guidelinetext ?term ?trialid
where {
?s ctec:hasText ?guidelinetext.
?s1 sct:hasRelations ?s.
?s1 sct:hasGuidelineItemID ?guidelineid.
?s ctec:hasFragment ?e1.
```

```

?e1 ctec:hasTerm ?term.
?e1 ctec:hasCUI ?conceptid.
?e2 ctec:hasCUI ?conceptid.
FILTER (!(?e1=?e2)).
?e1 ctec:hasTerm ?term.
?e3 ctec:hasObject ?e2.
?s2 ctec:hasFragment ?e3.
?s2 sct:NCTID ?trialid.}

```

One of the answers of this query is that the guideline with guidelineid 1002-zsh140412_4_1, with the guidelinetext “Adding MRI to mammography for the screening of high-risk women results in a higher sensitivity for breast cancer” is connected to the trial with trialID NCT00112749 by the term “screening”.

The last example shows that we search for the information of the evidence level over guidelines, like this:

```

select distinct ?conceptid ?guidelineid ?guidelinetext ?evidenceLevel
where {
?e1 ctec:hasTerm "flat epithelial atypia".
?e1 ctec:hasCUI ?conceptid.
?e2 ctec:hasObject ?e1.
?s ctec:includes ?e2.
?s ctec:hasText ?guidelinetext.
?s1 sct:hasRelations ?s.
?s1 sct:hasGuidelineItemID ?guidelineid.
?s1 sct:evidenceLevel ?evidenceLevel.
FILTER (?evidenceLevel <= 2)
}

```

Those example queries show the feasibility of reasoning based on the proposed lightweight formalization of evidence-based guidelines. The results encourage us to perform real experiments in collaboration with medical experts in the near future.

6 Discussion and Conclusions

6.1 Future Work

The advantage of the NLP tool is that it provides not only the concept annotation, but also relation statements with the RDF NTriple format. That would be quite convenient for us to check the similarity over fine-grained structures over two statements. One of the application scenarios of those semantic similarity and relevance checking is the guideline update, a task in the EURECA project.

Guideline update concerns how to transfer new research findings into existing clinical guidelines (e.g. the national breast cancer guidelines). These research findings can be stronger evidence or even new evidence for update of existing guidelines. It can have an effect on the care decision making based on the guideline.

Those new research findings can be collected from the latest publications, like those papers in PubMed, reports from clinical trials, or other available sources.

As one of the future work of this paper, we are going to develop a method to incorporate those new research results are linked to the interface of clinical guidelines on the SemanticCT system and the EURECA platform for the users (i.e., Guideline developers). With the help of the semantic representation of evidence-based clinical guidelines, we are going to create a model to check the relevance of new research results to the chosen clinical guidelines and demonstrate that the increased level of evidences brings substantial benefits to the clinical decision support applications.

6.2 Related Work

There has been a large body of work on providing formal representations of medical guidelines, using a wide variety of representation languages, such as the Arden Syntax, PROforma [6,7], Asbru [14], EON [16] and GLIF [12,13].

These earlier approach differ from the work presented in this paper in two important ways: the “semantic weight” of the representation, and the degree of automation of the modelling process (and these two are in fact coupled), as follows:

Most of the existing modelling languages are “semantically heavyweight”: they try to capture as much as possible of content of the clinical guideline, including control structure, applicability conditions, intentions, etc. As a consequence, these languages are very rich, with many features and high expressivity.

Consequently, the process of modelling guidelines in these languages is inevitably a manual task. As example, we take the pioneering work on the Digital Electronic Guideline Library (DeGeL) [15]. It facilitates gradual conversion of clinical guidelines from text to a formal representation in a chosen guideline ontology. The architecture supports use cases like guideline classification, semantic markup, context-sensitive search, browsing, run-time application, and retrospective quality assessment. Similar observations could be made for toolkits that support Asbru, ProForma, etc.

In contrast, in our case we are using NLP techniques that result in a light weight formalisation in the form of annotations. This means that we do not capture the full semantics of the guideline, but as we have shown in the previous section, this lightweight semantics is still sufficient to support a number of relevant and non-trivial use cases (and in fact, use cases which are not immediately supported by using the existing languages and environments).

The main difference is perhaps that our representation is specifically geared towards catching the evidence levels of guideline recommendations, an aspect that is missing from some of the older guideline representation languages.

GEM cutter is an XML guideline editor which is developed for GEM (the Guideline Elements Model), an XML-based guideline document model [9]. It can also serve as a tool for the generation of the XML data for evidence-based clinical guidelines. However, the XML data are just intermediate results of the

semantic representation of guidelines. The target in our approach is to obtain the semantic data so that they can be loaded into a triple store.

6.3 Conclusions

In this paper, we have proposed a framework of semantic representation of evidence-based clinical guidelines by using the Semantic Web standards, such as RDF/RDFS/OWL. We have reported how to use the XMedlan NLP tool to generate semantic data of evidence-based clinical guidelines. We have shown several example queries of the lightweight formalism of evidence-based clinical guidelines on the semantic operability. The relation extraction of the guideline statements provides an approach for semantic similarity and relevance checking between guideline statements and the statements in PubMed, which can be used for the use case of guideline update in the future.

Acknowledgments. This work is partially supported by the European Commission under the 7th framework programme EURECA Project (FP7-ICT-2011-7, Grant 288048).

References

1. Ait-Mokhtar, S., Bruijn, B.D., Hagège, C., Rupi, P.: Initial prototype for relation identification between concepts, D3.2. Technical report, EURECA Project (2013)
2. Ait-Mokhtar, S., Chanod, J.-P., Roux, C.: Robustness beyond shallowness: incremental deep parsing. *Nat. Lang. Eng.* **8**(2), 121–144 (2002)
3. Claerhout, B., De Schepper K., et al.: Initial EURECA architecture, D2.2. Technical report, EURECA Project (2013)
4. de Clercq, P., Blom, J., Korsten, H., Hasman, A.: Approaches for creating computer-interpretable guidelines that facilitate decision support. *Artif. Intell. Med.* **31**, 1–27 (2004)
5. Fensel, D., van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Della Valle, E., Fischer, F., Huang, Z., Kiryakov, A., Lee, T., School, L., Tresp, V., Wesner, S., Witbrock, M., Zhong, N.: Towards LarKC: a platform for web-scale reasoning. In: Proceedings of the IEEE International Conference on Semantic Computing (ICSC2008). IEEE Computer Society Press, CA, USA (2008)
6. Fox, J., Johns, N., Lyons, C., Rahmzadeh, A., Thomson, R., Wilson, P.: Proforma: a general technology for clinical decision support systems. *Comput. Methods Programs Biomed.* **54**, 59–67 (1997)
7. Fox, J., Johns, N., Rahmzadeh, A., Thomson, R.: Proforma, approaches for creating computer-interpretable guidelines that facilitate decision support. In: Proceedings of Medical Informatics Europe, Amsterdam (1996)
8. Huang, Z., ten Teije, A., van Harmelen, F.: SemanticCT: a semantically-enabled system for clinical trials. In: Riaño, D., Lenz, R., Miksch, S., Peleg, M., Reichert, M., ten Teije, A. (eds.) KGC 2013 and ProHealth 2013. LNCS, vol. 8268, pp. 11–25. Springer, Heidelberg (2013)
9. Karras, B., Nath, S., Shiffman, R.: A preliminary evaluation of guideline content markup using GEM - an XML guideline elements model. In: Proceedings of AMIA Annual Symposium (2000)

10. Lindberg, D., Humphreys, B., McCray, A.: The unified medical language system. *Methods Inf. Med.* **32**(4), 281–291 (1993)
11. NABON.: Breast cancer, dutch guideline, version 2.0. Technical report, Integraal kankercentrum Netherland, Nationaal Borstkanker Overleg Nederland (2012)
12. Peleg, M., Boxwala, A., Ogunyemi, O., et al.: GLIF3: The evolution of a guideline representation format. In: *Proceedings of AMIA Annual Fall Symposium*, pp. 645–649 (2000)
13. Peleg, M., Boxwala, A., Tu, S., Ogunyemi, O., Zeng, Q., Wang, D.: Guideline interchange format 3.4. Technical report (2001)
14. Shahar, Y., Miksch, S., Johnson, P.: The asgaard project: a task specific framework for the application and critiquing of time oriented clinical guidelines. *Artif. Intell. Med.* **14**, 29–51 (1998)
15. Shahar, Y., Young, O., Shalom, E., Mayaffit, A., Moskovitch, R., Hessing, A., Galperin, M.: DEGEL: a hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines. In: Dojat, M., Keravnou, E., Barahona, P. (eds.) *AIME 2003. LNCS (LNAI)*, vol. 2780, pp. 122–131. Springer, Heidelberg (2003)
16. Tu, S., Musen, M.: A flexible approach to guideline modeling. In: *Proceeding of 1999 AMIA Symposium*, pp. 420–424 (1999)
17. Witbrock, M., Fortuna, B., Bradesko, L., Kerrigan, M., Bishop, B., van Harmelen, F., ten Teije, A., Oren, E., Momtchev, V., Tenschert, A., Cheptsov, A., Roller, S., Gallizo, G.: D5.3.1 - requirements analysis and report on lessons learned during prototyping. LarKC project deliverable, June 2009