Using Semantic Web technologies for Clinical Trial Recruitment

Paolo Besana[1], Marc Cuggia[1], Oussama Zekri[2], Annabel Bourde[1],Anita Burgun[1]

[1] Université de Rennes 1, [2] Centre Eugène Marquis

Abstract. Clinical trials are fundamental for medical science: they provide the evaluation for new treatments and new diagnostic approaches. One of the most difficult parts of clinical trials is the recruitment of patients: many trials fail due to lack of participants. Recruitment is done by matching the eligibility criteria of trials to patient conditions. This is usually done manually, but both the large number of active trials and the lack of time available for matching keep the recruitment ratio low.
In this paper we present a method, entirely based on standard semantic web technologies and tool, that allows the automatic recruitment of a patient to the available clinical trials. We use a domain specific ontology to represent data from patients' health records and we use SWRL to verify the eligibility of patients to clinical trials.

1 Introduction

Clinical trials are the gold standard for testing therapies or new diagnostic techniques that may improve clinical care. Patients are enrolled to clinical trials if they match the eligibility criteria that define the trials. The recruitment process is a particular point of weakness for clinical trials. The development of information technology in medicine, and particularly in hospitals, offers a good opportunity to support and improve the recruitment process.

The work presented in this paper aims at suggesting the clinical trials to which a patient could be enrolled. It fits into the standard procedure, mandated by the French national oncology guideline, of evaluating cancer patients in multidisciplinary meetings. Doctors from different disciplines meet periodically to discuss and decide the treatment of patients. Clinical trials are about treatments, and the decision of enrolling a patient is taken during these meetings.

The system, by finding clinical trials in which a patient may be enrolled, takes a different perspective to clinical trial recruitment, usually oriented towards finding patients for a clinical trial. It is centred on patients, and it is possibly more acceptable by doctors, as it does not interrupt their workflow. Our project analysis and evaluation is based on the clinical trials and patients' data discussed in the multidisciplinary meetings in the Centre Hospitalier Universitaire of Rennes (France), between September and December 2009.

Matching trials to patients requires the formalisation of patient conditions and of eligibility criteria in such a way that their correspondences can be computed and found. Patients' data and criteria have usually a different level of

abstraction. Data are specific and precise (for example, Prostate Ductal Adenocarcinoma, Cribriform Pattern), while criteria need to include cases that are different within determined boundaries (for example, Invasive Prostate Carcinoma).

The fundamental hypothesis we make in the project is that mapping terms from patients' records and eligibility criteria to a formal ontology both minimises the risk of ambiguity and allows automated reasoning. The medical domain, and oncology in particular, has a wealth of well-established ontologies that can be used, which were originally developed as terminology services for uniquely identifying diseases, symptoms, and therapies. Additionally, ontologies written in OWL have clearly defined expressivity and computational properties and they can also exploit tools and applications both for authoring (such as Protégé) and for reasoning (such as Jena, Pellet, Fact++).

The work presented in this paper focuses on the use of OWL and SWRL for representing patients' data and eligibility criteria, and reason about them. The goal is to show that it is possible to identify a workable formalism within the boundaries of Description Logics that can be used for the whole process of recruitment, without the need to add external resources.

2 Clinical Trials

Clinical trials (CTs) are fundamental for evaluating therapies or new diagnosis techniques. They are the most common research studies designed to test the safety and/or the effectiveness of interventions. A CT may address issues such as prevention, screening, diagnosis, treatment, quality of life or genetics, and each trial is designed to answer specific scientific questions. CTs are based on statistical tests and population sampling, and because they rely on adequate sample sizes it is common for CTs to fail in their objectives because of the difficulty of meeting the necessary recruitment targets in an effective time and at reasonable cost.

The two most important issues that must be decided early in the design of a CT are the population of interest (which determines the eligibility criteria of the trial) and the sample size required to give sufficient statistical power for analysis. Reduced sample size reduces the power of the study, but relaxing eligibility criteria to allow a larger population of interest (and hence a larger pool from which to recruit) introduces a confounding element where factors that are not the prime focus of the study cannot be excluded.

Patient data, either acquired during the clinical care process or contained in Electronic Health Records (EHR), could be reused to automatically apply eligibility criteria

The features of the population of interest for a clinical trial are defined by the eligibility criteria of the trial. These characteristics determine the rules to be applied for building the sample of subjects. They may include age, gender, medical history and current health status. Eligibility criteria for treatment studies often require that patients have a particular type and are at a particular stage of their disease.

Enrolling participants with similar characteristics helps to ensure that the results of the trial will be due to what is under study and not other factors. A second function of eligibility criteria is to exclude patients who are likely to be put at risk by the study, minimizing the risk of a subject's condition worsening through participation.

3 Problem Description

The goal of the system presented in this paper is to select the clinical trials in which a patient might be enrolled, among those currently active in a hospital. The list of selected clinical trials are then evaluated by the doctors in the multi-disciplinary meeting. In particular, it is important to remove trials that are either not relevant or with mismatch conditions, in order to provide the physicians with a list of focussed suggestions.

The suggestions are based on the available information at the moment of the meeting, which can be weeks before the trial actually starts. Criteria referring to the patient conditions at the moment of the trial cannot be considered in this stage and are discarded: the project focuses on a subset of the criteria, called pre-screening criteria. The system is applied to clinical trials concerning prostate cancer.

There are three main aspects to be considered: how to represent patients' data in a format that can be queried; how to represent the eligibility criteria; and how to match criteria to patients, dealing with the difference in abstraction discussed in Section 1. In this paper we present an approach that uses only OWL and SWRL to represent data and criteria, and to reason. It can be considered a low-level representation: directly computable, but not for human use. However, it is possible to find representations at a slightly higher level that can be directly converted into SWRL using re-write rules.

3.1 Patient Data

Results of exams are stored in patients' records. While historically kept in physical folders, they are beginning to be stored in an Electronic Health Records (EHR), often in natural language, or as scanned images. There is an ongoing effort to formalise their representation, in order to simplify search, and to allow interoperability between different systems.

In our project, the data is currently in free text, but an expert operator will convert it into a more formal model before running the matching process.

Patients' data are represented according to an Information Model, such as the HL7 Reference Information Model (RIM). An information model defines what is the information that needs to be collected and gives it a semantics. It is often a set of classes (also called templates) with the attributes and methods (the class patient, for example). The HL7 RIM contains 5000 attributes to cover terminology requirements. Semantics is provided by a reference ontology.

date of birth: 11 October 1935
Relevant elements:
diagnosis: prostate adenocarcinoma
June 2007: radical prostatectomy
Gleason 6 = 3+3, pT3a R0
Initial PSA (Prostate Specific Antigene) marker = 9.48
one month after surgery=undetectable
after 12 months=0.26 ng/ml

Fig. 1. Example of patient record available at multidisciplinary meeting

3.2 Eligibility Criteria

The eligibility criteria consist of a set of inclusion criteria defining the characteristics mandatory in the population of interest and a set of exclusion criteria defining the characteristics to be avoided. Usually the negation of an exclusion criterion becomes an inclusion criterion and viceversa.

Eligibility criteria can be simple, stating the value of a single observable entity, such as the diagnosis (diagnosis of prostate adenocarcinoma), or can be qualified by other properties (diagnosis of prostate adenocarcinoma, confirmed by histology). The values can be at different level of abstraction: for example, the diagnosis can specify a particular type of cancer, or can be more generic and include different forms of cancer, possibly by specifying only the location (diagnosis of prostate neoplasm) or some features of the cancer (invasive cancer).

Criteria can define the acceptable value of some medical parameters (Prostate Specific Antigene PSA $>$ 5ng/l), of some personal attribute of the patient (age $>$ 18, age $<$ 75) or can specify the staging of the disease (such as the Classification of Malignant Tumours, TMN, or the Karnofsky score and the Zubrod score, used to measure the general patient's well-being). There can be alternative values (patient in stage pT2, pT3 or pT4). The staging system used in the patients data may be different from the one used in the criteria. Conversion tables can address this issue.

The criteria can specify time constraints that refer to other events: PSA value $>$ 9 ng/l 6 months after surgery, or PSA $<$ 2 a month before inclusion.

A criterion can contain a conjunction of other criteria (diagnosis of X and grade pT2), or can be the disjunction of criteria (PSA $>$ 9 or grade pT2).

Inclusion criteria are usually positive expression, while exclusion criteria are introduced by negation ("no ...", "absence of..."). Some composite criteria may contain both positive and negative statements (invasive cancers, excluding skin cancer).

Figure 2 shows an example of the eligibility criteria used for selecting patients in clinical trial for a therapy for prostate cancer run in the university hospital of Rennes in 2009.

4 State of the Art

Inclusion:
1) Histologically proven cancer localised in the prostate
2) Absence of metastases
3) Cancer in intermediate prognostic group : - T2a <= T < T3a - or T1b/c with PSA >= 10ng/ml - or T1b/c with Gleason score >= 7
4) PSA < 30ng/ml with a normal calibration value of 4ng/ml
5) age < 77 years
6) life expectancy >= 10 years
7) OMS-WHO=ZUBROD <= 1 [Zubrod]
Exclusion:
1) history of invasive cancer unless it is older than 5 years
3) PSA >=30ng/ml in two successive measurement (even if the latest is lower than 30)
4) history of pelvic radiotherapy
5) history of radical anterior prostectomy due to cancer
6) previous hormonotherapy or castration

Fig. 2. Eligibility criteria for a clinical trial for an adjuvant therapy for prostate cancer

A detailed and extensive overview of the formalisms used for representing eligibility criteria is given in [10]. In the paper the authors distinguish different types of expression languages for eligibility criteria. We summarise three categories, which include most of the projects.

ad hoc expression normally driven by use cases more than by theoretical basis. Ad hoc languages define a set of parameters that can take boolean, numeric or enumerated values. The languages provide comparison and logical operators. Some ad hoc languages are based on a rich information model, such as the HL7 RIM. In general they have a limited capability of using formal reasoning methods such as temporal constraints or predicate logic. However, ad hoc languages proved very popular, and are used in various projects.

Arden Syntax is a hybrid between a production rule system and a procedural formalism. It has been chosen as standard for HL7. It provides rich time functions and explicit links to clinical data embedded in curly brackets. It is more expressive than most ad hoc languages. It lacks declarative properties and defined semantics for temporal comparison and data abstraction. It is well supported.

logic based languages vary in expressivity. Systems overviewed include one based on SQL (based on relational algebra), one on Protege Constraint language (PAL) and one on Description Logic [9].

Because of its relation with the project presented in the paper, we will describe [9] in more detail. This work aims at demonstrating how it is possible to use ontologies for reasoning in health informatics. The authors use the problem of matching eligibility criteria to patients' condition as case study. They

focus on the problems of knowledge engineering and of scalability. They analyse the mapping of the representation of patients' data used in a hospital, based on a local terminology, to a formal medical ontology - SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms)[1]. The problem of scalability is addressed using SHER, an OWL-DL reasoner developed by IBM for dealing with large ontologies. Patients' data are observations that connect instances of SNOMED CT classes. Eligibility criteria are class definitions that are used to query observations via subsumption.

In the Epoch project [7] the researchers have developed a framework for managing the overall process of clinical trials. They created a suite of OWL ontologies covering the different phases of the process, but they do not provide a detailed explanation about the representation of patients' data and clinical trials.

5 Method

As we have seen in Section 3, formalising and reasoning over eligibility criteria requires the possibility of reasoning over ontologies when criteria are expressed as generic conditions, or when only some attributes of the diseases are specified,. It also needs to be able to represent and reason over data types and over time. It should allow the composition of criteria both through disjunction and conjunction. An additional requirement is the traceability of the results: if a patient is selected or rejected for a clinical trial, it must be possible to identify the observations supporting the criteria.

As stated in Section 1, we make the fundamental hypothesis that mapping terms from patients' records and eligibility criteria to a formal ontology minimises the risk of ambiguity and allows automatic reasoning. The ontology plays both the role of reference terminology and of background knowledge for reasoning.

The system needs to find the clinical trials to which a patient might be enrolled. Before a multidisciplinary meeting, all the active Trials of a hospital are loaded into the system. Then each of the patients to be discussed in the meeting is loaded into the system, one at the time. The eligibility criteria are matched to patient's observations. The criteria are then aggregated and a list of clinical trials with satisfied inclusion criteria and without satisfied exclusion criteria is extracted.

The system exclusively uses OWL and SWRL to represent patient data and eligibility criteria: everything is loaded into an ontology, and all the operations take place within the ontology.

5.1 Choosing the Medical Ontology

The first step is to identify the ontology that can be used for these roles. For its role as reference terminology, an ontology needs to provide a good coverage of the terms appearing in patients' records and in criteria: it should be possible

[1] http://www.ihtsdo.org/snomed-ct/

to map terms in the text to entities already defined in the ontology, or to easily define new entities, using other entities and the compositional grammar of the ontology. Following the criteria presented in [4], we also evaluated availability (open source or licensed) and its format.

In order to evaluate coverage an expert in clinical trial selected 200 criteria from `clinicaltrials.gov`. We used MetaMap [2] to map the extracted criteria, written in free text, to concepts in the UMLS metathesaurus [3]. UMLS (Unified Medical Language System) connects terms from over 100 medical terminologies: each concept in UMLS is mapped to different terminologies. We checked, for each UMLS concept found by MetaMap, whether there was a corresponding term in the different terminologies. From the evaluation, it resulted that NCI Thesaurus (NCI-T) [8] provides the best coverage with 75%. NCI-T contains 75000 classes and it is developed specifically for oncology by the National Cancer Institute, US. NCI-T is followed by SNOMED CT, a large ontology (over a million classes) covering all medical domains, with a coverage of 65%.

Regarding availability, NCI-T is open source, while SNOMED CT has a very expensive licence. Both are written in a Description Logic: SNOMED is $\mathcal{ER}$++, NCI Thesaurus is $\mathcal{SH}(D)$. However, NCI-T is directly available in OWL1.1 [6], while SNOMED is distributed in database tables and requires conversion . NCI-T introduces a particular idiom, intended for human use, for properties. Some of its properties have the prefix `may_have`, while some of the others have the prefix `excludes_`. The `may_have` properties are used when a class of diseases have subclasses that may or may not present a particular feature. A `may_have_`[*feature*] property has a corresponding subproperty `has_`[*feature*] that is used in subclasses that have the feature. The `excludes_[`*feature*`]` property is used to specify that a particular disease does not present a feature or a symptom. These properties are intended for human use: from a Description Logics perspective these properties have no particular meaning, and it is not possible to use them to verify consistency.

At the end of the evaluation process, we opted for NCI-Thesaurus as background knowledge. Additionally, the use of a domain-specific ontology requires a system that should be minimally coupled with the ontology itself, to allow portability in other domains.

The support for data types in OWL1.1 is not particularly powerful, as it is not possible to specify ranges. Similarly the reasoning over time is hard in OWL. The use of SWRL (Semantic Web Rule Language) allows us to overcome these limitations: in SWRL it is possible to write rules that state ontological properties and contain built-in functions. In Protege 3.4, the available built-ins include numerical comparators (greaterThan, lessThan,...) and temporal comparators.

The core of the system is a simple ontology, that glues together the components: it imports NCI-T ontology and the ontologies that define SWRL built-ins and possibly the classes they require. This glue ontology defines the classes that are used to represent patient data and the eligibility criteria in clinical trials.

Figure 3 shows the general architecture: TBox contains the glue ontology that imports NCI-T and the SWRL ontologies. The SWRL rules are definitions,

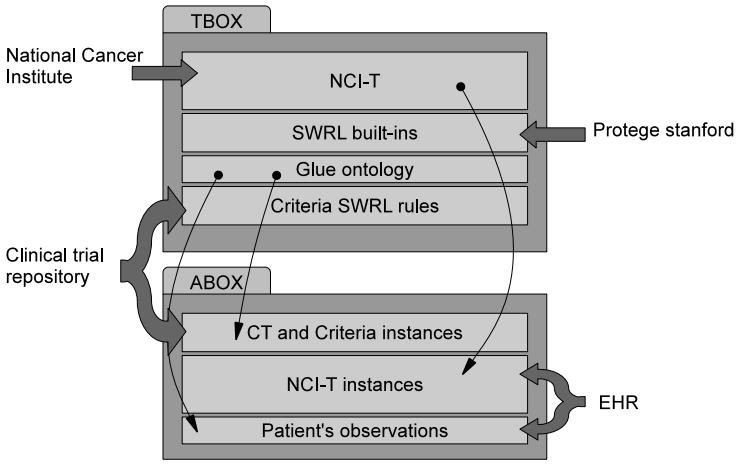

Fig. 3. General architecture of the system

and therefore are part of the TBox. The ABox contains the instances of NCI-T classes, that are linked by observations, and the instances of the criteria.

5.2 Representing Patient Data

Medical ontologies like SNOMED CT or NCI Thesaurus were originally developed as terminology services, and were not conceived for representing patients's data directly. The information about patients is captured by the Information Model. Some of the Information Models, such as the HL7 v3.0, use an ontology as reference for its terms. In order to allow portability, we use a thin information layer composed by observations that connect values taken from the real information model.

In our system, an observation is the tuple:

$$\langle observable\_entity, observable\_property, observed\_value \rangle$$

It is a reified relation connecting an *observable entity* (a measurement, a medical finding, the result of an exam), an *observable property* (value, numeric value, date, method, ...), and the *observed value* of the property (it can be an instance of a class, or a datatype). In Owl, it is represented as an instance of the class Observation. All entities and values, unless they are datatypes, need to be instances: OWL-DL does not support the relations between classes (it would make it into OWL-FULL which is undecidable). If the class definition contains as sufficient condition properties with existential restriction, then an instance for the restricted class is created. To avoid a cascading effect of having to create instances for all the properties of the instances created to fill properties, only the properties of instances created directly from the observation are filled. Filling the restricted properties is required for reasoning later: SWRL reasons on instances only. If a disease class has the restriction `has_lesion some Invasive_Lesion`, the instance will be correctly classified as an invasive disease only if its property `has_lesion` is set to an instance of `Invasive_Lesion`. This applies only to the existential restrictions: the universal restriction means that either the property has no value, or if it has a value it can only be an instance of the class in the

restriction. A particular symptom or effect either is present or it is not. As we stated above, NCI-T use properties with prefixes `may_have_` and `exclude_`. Because of its intended meaning, the restriction applied to `may_have_` properties can be exclusively be of universal type, and so we do not need to consider them.

Example

We present here the translation into OWL of the observation shown in Figure 1. First of all, a set of predefined instances is loaded into the ontology, used in all the observations:

`valuep` instance of `ncit:Value;`
`numvaluep` of `ncit:NumericValue;`
`datep` of `ncit:Date; procp` of `ncit:Procedure:;`
`now` of `temporal:ValidInstant`

Date of Birth: 10 November 1935

The observable entity is an instance of NCI class `ncit:Birthday`, the observable property is an instance of the class `ncit:Value` (instantiated as `valuep` in the previous step) and the observed value is represented as an instance of the class `validInstant` from the temporal ontology,

```
ncit:BirthDay:birth_day_1;
Temporal:validInstant:i1:[has_time:10November1935];
Observation:o1:[has_observable=Birth_Day_1,
  has_observable_property=valuep,has_value:i1];
```

diagnosis: prostate adenocarcinoma

The observation links the instances of the class `ncit:Diagnosis` and of the class `ncit:Prostate_Adenocarcinoma`

```
ncit:Diagnosis:diagnosis1;ncit:Prostate_Adenocarcinoma:ac1;
Observation:o2:[has_observable=diagnosis1,
  has_observable_property=valuep,has_value=ac1];
```

june 2007: radical prostatectomy

We need to observations: one to cover the value of Clinical procedure, the other to cover its date

```
ncit:Clinical_Procedure:cp1;  ncit:Radical_Prostatectomy:rp1;
Temporal:validInstant:i3:[has_time:1July2007];
Observation:o3:[has_observable=cp1,
  has_observable_property=valuep,has_value=rp1];
Observation:o4:[has_observable=cp1,
  has_observable_property=datep, has_value=i3];
```

PSA after 12 months=0.26 ng/ml

We need one observation for the numerical value of the PSA and one for the date of the exam

```
ncit:PSA_Assay:psa1;   Temporal:Instant:i6:[has_time:1July2008];
Observation:o7:[has_observable=psa2,
  has_observable_property=numvalp, has_numeric_value=0.26];
Observation:o8:[has_observable=psa2,
  has_observable_property=datep, has_ value=i6];
```

5.3 Representing Clinical Trials

Criteria are queries over the observations containing patients' data. Patients whose data match all the inclusion criteria are included in the clinical trial, unless they match an exclusion criteria.

SWRL [5] provides a high-level syntax for horn-like rules: a SWRL rule has the form of an implication with an antecedent (body) and a consequent (head): if the antecedent holds, the condition specified in the consequent holds. SWRL maintains the expressivity of OWL-DL, with a set of additional features such as built-in functions for data types. SWRL is monotonic: it is not possible to retract or change what is already assessed, but only to add something new. OWL and consequently SWRL rely on the Open World Assumption: the lack of some information is considered ignorance. On the contrary, the Closed World Assumption, used in most of the other programming languages and in databases, considers to be false that which is not known to be true . With the Open World Assumption, it is not possible to verify whether something is false unless it is explicitly stated as false. It is a realistic assumption in the medical domain: it may not be possible to collect all the information, both because in different phases different information is available, and because some exams are probabilistic (for example, the presence of metastases cannot be completely ruled out if some samples give negative results).

In Protégé 3.4 SWRL rules are computed using Jess, a proprietary library based on the RETE algorithm and developed by the Sandia Laboratories . The ontology and rules need to be converted to Jess (simple with Protégé), and then the engine is run. However, a limitation of Jess is its lack of support for inferred classifications in the ontology. Protege 4.1 uses only OWL reasoners for SWRL, and therefore it supports inferred classifications. We started the development before the availability of version 4.1,. However, our system can work on both versions of Protégé.

Because of SWRL monotonicity, we can only add new information: every satisfied rule adds the observations it matches to the support list of a criterion. Supported inclusion criteria are added as supporting arguments to a clinical trial, while supported exclusion criterion are added as arguments against the clinical trial. If a clinical trial has arguments against it, it is excluded from the list of possible trial for a patient.

In our system, criteria are instances of Inclusion or Exclusion Criteria classes, defined in the glue ontology. Each criterion has a set of rules: if one matches any of the observations, it adds the observation to the list of support of the criterion. After running all the criteria rules, the criteria are aggregated either in favour or against the clinical trials

Criteria also can be annotated with a human-readable description, to facilitate the final report. The left side of the rule represents the condition, and must match one or more observations. The right hand side adds the matching observations to the list of supporting observations:

$\texttt{Observation}(?\texttt{o}) \wedge \texttt{C1} \wedge \texttt{C2}... \rightarrow \texttt{is\_supported}(\texttt{Pid}, ?\texttt{o})$

where Pid is the identifier of the criterion supported by the observation. Criteria that verify the value of an observable entity can be represented directly:

$$\texttt{Observation}(?o) \wedge \texttt{has\_observable}(?o, \texttt{Entity}) \wedge \\ \texttt{has\_value}(?o, ?V) \wedge \texttt{ExpectedEntity}(?V) \\ \rightarrow \texttt{is\_supported}(\texttt{Pid}, ?o)$$

Similarly, for criteria about the numeric value of an observable entity:

$$\texttt{Observation}(?o) \wedge \texttt{has\_observable}(?o, \texttt{Entity}) \wedge \\ \texttt{has\_numeric\_value}(?o, ?V) \wedge \texttt{swrlb:comparator}(?V, \texttt{Num}) \\ \rightarrow \texttt{is\_supported}(\texttt{Pid}, ?o)$$

Where the swrlb:comparator is one of the numeric built-in functions (such as greaterThan, lessThan, greaterOrEqual, lessOrEqual) and Num is constant. When an observable can have more than a value, it is necessary to create a rule per value, all supporting the same criterion.

When the criteria have temporal requirements, the rule needs to extract the observation about the date relative to the observed entity (if available), and use the temporal built-in operators of SWRL. The glue ontology must therefore import the Temporal Ontology[2] that defines temporal operations, based on Allen's temporal logic [1], and some basic classes used in defining time intervals. Some of the criteria refer to dates that are implicit (like the current time, inserted before running, as we saw above).

When the criteria refer to partially defined concepts that do not have a corresponding entity in the background ontology the clean solution would be to define a new class with sufficient conditions and push it in the TBox before loading the rules. However, as said above, Jess, the engine used by Protege 3.4 for SWRL does not support inferred relations. Equivalently, the sufficient conditions are specified in SWRL.

The work presented in this paper focuses on the use of OWL and SWRL for representing patients' data, eligibility criteria and reasoning about them. The goal is to show that it is possible to identify a workable formalism within the boundaries of DL. While the representation is computable, and results can be obtained, it is awkward for human operators. Most of the criteria fit in a relatively limited set of patterns. The criteria can be inserted using these patterns, and then SWRL rules are generated using rewrite rules.

Example.

We present as example the eligibility criteria shown in Figure 2. We first create an instance for the clinical trial:
ClinicalTrial:ct1
We then show the representation of a subset of the criteria.
Cancer localised in the prostate

[2] http://swrl.stanford.edu/ontologies/built-ins/3.3/temporal.owl

In this case, we need to exploit the subsumption mechanism and the multi-hierarchy nature of NCI-T. The observable entity is of type ncit:Diagnosis, and the value needs to be an instance of both the class cancer and the class prostate disorder.

`InclusionCriterion:ic1:[CT=ct1]`

r1: $\texttt{Observation}(?o)\texttt{has\_observable}(?o, ?e) \wedge \texttt{ncit} : \texttt{Diagnosis}(?e) \wedge \texttt{has\_observable\_property}(?o, \texttt{valuep}) \wedge \texttt{has\_value}(?o, ?v) \wedge \texttt{ncit} : \texttt{Neoplasm}(?v) \wedge \texttt{ncit} : \texttt{Prostate\_Disorder}(?v) \rightarrow \texttt{supported\_by}(\texttt{ic1}, ?o)$

Absence of metastases

This criterion, presented as an inclusion criterion, needs to be treated as an exclusion criterion: the enrollment is excluded if there is a metastasis. We use the diagnosis to verify whether it is a metastatic disease. We could also use rules about the stage of cancer.

`ExclusionCriterion:ec1:[CT=ct1]`

r2: $\texttt{Observation}(?o) \wedge \texttt{has\_observable}(?o, ?e) \wedge \texttt{has\_observable\_property}(?o, \texttt{valuep}) \wedge \texttt{ncit} : \texttt{Disease\_has\_finding}(?e, ?f) \wedge \texttt{ncit} : \texttt{Metastatic\_Lesion}(?f) \rightarrow \texttt{supported\_by}(\texttt{ec1}, ?o)$

Cancer in intermediate prognostic group : - T2a <= T < T3a - or T1b/c with PSA >= 10ng/ml - or T1b/c with Gleason score >= 7

We split the criterion into 10 alternative rules (4 for the stages T2a-T3a, 2 for the T1b/c with the specified PSA value, and two for T1b/c with specified Gleason score), one of which has to be supported by an observation

`InclusionCriterion:ic2a:[CT=ct1]`

r3a: $\texttt{Observation}(?o) \wedge \texttt{has\_observable}(?o, ?e) \wedge \texttt{ncit} : \texttt{Finding}(?o, ?e) \wedge \texttt{has\_observable\_property}(?o, \texttt{valuep}) \wedge \texttt{has\_value}(?o, ?f) \wedge \texttt{ncit} : \texttt{pT2a\_Stage\_Finding}(?f) \rightarrow \texttt{supported\_by}(\texttt{ic2}, ?o)$

...

r3d: $\cdots \texttt{ncit} : \texttt{pT3a\_Stage\_Finding}(?f) \rightarrow \texttt{supported\_by}(\texttt{ic2}, ?o)$

The criterion is supported also if the finding is T1b/c with a value of PSA above 10ng/ml, or with a gleason score above 7

r3e: $\texttt{Observation}(?o1) \wedge \texttt{has\_observable}(?o1, ?e) \wedge \texttt{ncit} : \texttt{Finding}(?e) \wedge \texttt{has\_value}(?o1, ?f) \wedge \texttt{ncit} : \texttt{pT1b\_Stage\_Finding}(?f) : \wedge \texttt{Observation}(?o2) \wedge \texttt{has\_observable}(?o2, ?e2) \wedge \texttt{ncit} : \texttt{PSA\_Assay}(?e2) \wedge \texttt{has\_numeric\_value}(?o2, ?v) \wedge \texttt{swrlb} : \texttt{greaterOrEqual}(?v, 10) \rightarrow \texttt{supported\_by}(\texttt{ic2}, ?o)$

r3f: $\cdots \wedge \texttt{ncit} : \texttt{pT1c\_Stage\_Finding}(?f) \wedge \cdots \rightarrow \texttt{supported\_by}(\texttt{ic2}, ?o)$

The Gleason score follows the same principle

NO history of invasive cancer unless it is older than 5 years

In this case, the criterion would be false if there was an observation about an invasive cancer more recent than 5 years before screening. Therefore the exclusion criteria queries about such events, using the observable Personal_Medical_History, that can be used to report medical events in the past, and the Invasive_Malignant_-Neoplasm class from NCI-T. The Adenocarcinoma of the Prostate, seen above, is one of its subclasses. We need to verify two observations about the same ob-

servable (the medical history): one relative to its value, and the other about its date.

ExclusionCriteria:ec2:[CT=ct1]

r4: $\texttt{Observation}(?\texttt{o1}) \wedge \texttt{has\_observable}(?\texttt{o1}, ?\texttt{e}) \wedge \texttt{Personal\_Medical\_History}(?\texttt{e}) \wedge$
$\texttt{has\_observable\_property}(?\texttt{o1}, \texttt{valuep}) \wedge \texttt{has\_value}(?\texttt{o1}, ?\texttt{v}) \wedge$
$\texttt{ncit} : \texttt{Invasive\_Malignant\_Neoplasm}(?\texttt{v}) \wedge \texttt{Observation}(?\texttt{o2}) \wedge$
$\texttt{has\_observable}(?\texttt{o2}, ?\texttt{e}) \wedge \texttt{has\_observable\_property}(?\texttt{o2}, \texttt{datep}) \wedge$
$\texttt{has\_value}(?\texttt{o}, ?\texttt{d}) \wedge \texttt{temporal} : \texttt{duration}(?\texttt{p}, \texttt{now}, ?\texttt{d}) \wedge \texttt{swrlb} : \texttt{lessThan}(?\texttt{p}, 5)$
$\rightarrow \texttt{supported\_by}(\texttt{ec2}, ?\texttt{o})$

where the built-in function `temporal:duration(?p, now,?D, ‘‘Months’’)` instantiates `?p` with the length of the interval beween the date `?d` and the date `now`

5.4 Aggregating the Results

A clinical trial has a list of arguments in favour and one against enrollment of the patient. The list in favour is filled by inclusion criteria, while the list against by exclusion criteria.

Once the criteria rules are run, the instances of the criteria will have the property supported_by either filled with one or more observations, or empty. Inclusion criteria with the property filled are supported, and they in turn support the clinical trial. Exclusion criteria with the supported_by property filled are arguments against the enrolling of the patient to the clinical trial.

At this point, rules specific to clinical trials are run to aggregate the results of the criteria.All inclusion criteria of a clinical trial need to be verified: there must be, for each clinical trial, a rule stating that the conjunction of all inclusion criteria must have at least one supporting argument:

$\texttt{supported\_by}(\texttt{cid1}, ?\texttt{a}) \wedge \texttt{supported\_by}(\texttt{cid2}, ?\texttt{b}) \wedge ...$
$\rightarrow \texttt{is\_supported}(\texttt{CT1}, \texttt{true})$

Then a generic rule for the exclusion criteria is run:

$\texttt{ExclusionCriterion}(?\texttt{c}) \wedge \texttt{has\_ct}(?\texttt{c}, ?\texttt{ct}) \wedge \texttt{supported\_by}(?\texttt{c}, ?\texttt{a})$
$\rightarrow \texttt{argument\_against}(?\texttt{ct}, ?\texttt{c})$

It is also possible to trace the criteria that were against the enrollment to a criteria, and for each criterion it is possible to trace the observation that supports it.

However, extracting the clinical trials that are supported without arguments against it requires to reason with the closed world assumption: according to the open world assumption, the empty list of arguments against enrollment is considered ignorance, and cannot be used to infer that the there are no arguments against enrollment. The last step needs to be performed outside the ontology. An external program obtains the list of all clinical trials and selects those that are both supported and have no arguments against.

Table 1 shows the state of the criteria and of the clinical trials at the end of the execution on the example data and criteria.

criteria	description	CT	support
ic1	cancer localised in prostate	CT1	o2
ec1	absence of metastases	CT1	
ic3	intermediate prognostic group	CT1	o5

CT	favour	against
CT1	ic1,ic3	-
CT2	...	...

Table 1. Example of results at the end of the execution

6 Evaluation and Discussion

The aim of this work is to show how it is possible to represent eligibility criteria of clinical trials using SWRL on top of a large domain specific ontology such as NCI Thesaurus. The first step is to assess how well eligibility criteria can be represented. An expert in clinical trials selected 97 criteria from `clinicaltrials.gov`, that are particularly representative pre-screening criteria. We started from a larger set, but some were removed as they required information which is not available at pre-screening time.

Out of the 97, 92 could be fully represented using SWRL. The problematic ones were caused by the lack of the corresponding entity in NCI. 11 needed disjunction, 20 contained numerical comparison, 14 required some form of temporal reasoning. About a third contained queries over more than one observation.

We also extracted four real clinical trials active during 2009 in the University Hospital of Rennes, and we selected 129 patients that were examined and assessed for the trials during the same year. The four trials have 67 different eligibility criteria, some appearing in more than a trial. Overall, 7 could not be represented: all contained terms that cannot be mapped to corresponding entities in NCI (life expectancy, hip replacement, cardio-vascular pathology, neuro-pathology, hypertension, under tutelage). It is easy to notice how terms are missing when they come from a domain that differs from oncology, the domain of NCI. The criteria are separated into inclusion or exclusion.

Some of the criteria can be directly translated into SWRL - we have seen some examples above. Others require more thought, especially these which involve temporal reasoning.

In our project time is not a stringent requirement: the matching of the patients to the available clinical trials is done offline, before the multidisciplinary meeting. The slowest step in the overall procedure is loading NCI-T ontology: on a dual core machine, with 8Gb of memory, takes over 100 seconds and 2Gb of memory. Importing data into the ontology is nearly instantaneous: we load one patient at the time, and only the clinical trials currently active in the hospital are loaded. The next bottleneck is the conversion of the ontology and of the SWRL rules into Jess, operation that takes on average 10 seconds. The actual running of the engine takes less than a third of a second (but as we explained above, the inferred relations are not considered by Jess). Compared to [9] we use a much smaller ontology (SNOMED CT is over a million classes, while NCI-T is only 75000). Loading the background ontology is performed once. The criteria

and the patients observation need to be inserted for every patient and deleted at the end of the matching.

7 Future Work

We plan to proceed in three directions, one addressing the cause of failure in representing criteria, another one dealing with the new version of NCI-T that will soon released and finally a third studying portability to other domains.

The choice of NCI-T over SNOMED CT has a few advantages, among which the smaller size of the ontology. However, there is a trade-off between smaller size and possibility of representing all the criteria: the main cause of failure was shown to be the lack of the corresponding concept in NCI-T for terms in the criterion. We need to address the problem of entities not defined in NCI-T: we plan to study the feasibility of importing fragments of other medical ontologies to cover these gaps.

A new version of NCI-T is currently under development: it will be release in OWL2.0, and it will exploit the new features on datatypes available in OWL2.0. Once the new version will be released, we will move the system to OWL2.0.

Patients and trials studied in this project concerned prostate cancer only. We plan to assess the results applying the system to different types of cancers. Moving to a domain different from oncology requires either identifying another domain-specific ontology, or using SNOMED CT, possibly extracting a relevant portion of this large ontology.

8 Conclusion

Clinical trials are required for the evaluation of medical treatments. Their weakness lies in the difficulty of recruiting enough patients in order to make them statistically meanigful. In this paper we have presented an approach based on OWL and SWRL that addresses the problem of recruitment of patients.

The patients' data, extracted from the Electronic Health Record are converted into observations, that are reified relations linking observable entities, such as measurements, diagnoses, results of exams, to attributes, such as value, date or method. The eligibility criteria are SWRL rules that match observations.

The evaluation showed that it is possible to represent the great majority of criteria, and the difficulties raise when entity in the background ontology cannot be found for terms in the criteria. Compared to the work in [9], the approach based on SWRL allows the representation and reasoning over temporal constraints in the criteria. Compared to the Epoch framework [7], this work focusses on the representation of patient data and eligibility criteria using a domain specific ontology.

9 Acknowledgments

We are grateful to Olivier Dameron and Fiona McNeill for their valuable comments. The project was funded by the French National Agency of research (ANR) under the TECSAN program 2008.

References

1. J.F. Allen. An interval-based representation of temporal knowledge. In *proceedings of the 7th IJCAI*, pages 221–226, 1981.
2. AR Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
3. O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database Issue):D267, 2004.
4. O Bodenreider1 and A Burgun. Towards desiderata for an ontology of diseases for the annotation of biological datasets. In *International Conference on Biomedical Ontology 2009*, july 2009.
5. I. Horrocks, P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission 21 May 2004. *World Wide Web Consortium*, 2004.
6. N.F. Noy, S. de Coronado, H. Solbrig, G. Fragoso, F.W. Hartel, and M.A. Musen. Representing the NCI Thesaurus in OWL DL: Modeling tools help modeling languages. *Applied ontology*, 3(3):173–190, 2008.
7. R.D. Shankar, S.B. Martins, M.J. O'Connor, D.B. Parrish, and A.K. Das. Epoch: an ontological framework to support clinical trials management. In *Proceedings of the international workshop on Healthcare information and knowledge management*, page 32. ACM, 2006.
8. N. Sioutos, S. Coronado, M.W. Haber, F.W. Hartel, W.L. Shaiu, and L.W. Wright. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43, 2007.
9. Kavitha Srinivas, Chintan Patel, James Cimino, Li Ma, Julian Dolby, Achille Fokoue, Aditya Kalyanpur, Aaron Kershenbaum, and Edith Schonberg. Matching patient records to clinical trials using ontologies. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*, volume 4825 of *LNCS*, pages 809–822, Berlin, Heidelberg, November 2007. Springer Verlag.
10. C. Weng, S.W. Tu, I. Sim, and R. Richesson. Methodological Review: Formal representation of eligibility criteria: A literature review. *Journal of Biomedical Informatics*, 43(3):451–467, 2010.