SemanticCT: A Semantically-Enabled System for Clinical Trials

Huang, Z.; ten Teije, A.C.M.; van Harmelen, F.A.H.

published in
Lecture Notes in Computer Science
2013

DOI (link to publisher)
10.1007/978-3-319-03916-9_2

document version
Early version, also known as pre-print

SemanticCT: A Semantically-Enabled System for Clinical Trials

Zhisheng Huang, Annette ten Teije, and Frank van Harmelen

Department of Computer Science,
VU University Amsterdam, The Netherlands
{huang,annette,Frank.van.Harmelen}@cs.vu.nl

Abstract. In this paper, we propose an approach of semantically enabled systems for clinical trials. The goals are not only to achieve the 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. We have implemented the proposed approach in a system called *SemanticCT*. SemanticCT is built on the top of LarKC (Large Knowledge Collider), a platform for scalable semantic data processing. SemanticCT has been integrated with large-scale trial data and patient data, and provided various automatic services for clinical trials, which include automatic patient recruitment service (i.e., identifying eligible patients for a trial) and trial finding service (i.e., finding suitable trials for a patient).

1 Introduction

Clinical trials provide tests which generate safety and efficacy data for health interventions. Clinical trials usually involve large-scale and heterogeneous data. The lack of integration and of semantic interoperability among the systems of clinical trials and the systems of patient data, i.e. electronic health record (EHRs) and clinical medical records (CMRs), is the main source of inefficiency of clinical trial systems. Thus, many procedures in clinical trials, such as patient recruitment (i.e., identifying eligible patients for a trial) and trial finding (i.e., finding suitable trials for a patient), have been considered to be laborious.

Enhancing clinical trial systems with semantic technology to achieve the semantic interoperability of large-scale and heterogeneous data would improve the performance of clinical trials significantly. Those semantically-enabled systems would achieve efficient and effective reasoning and data processing services in various settings of clinical trials systems.

In this paper, we propose an approach of semantically enabled systems for clinical trials. The proposed approach has been implemented in the system called *SemanticCT*[1]. The system provides semantic integration of various data in clinical trials. The system is designed to be a semantically enabled system of decision

[1] http://wasp.cs.vu.nl/sct

support for various scenarios in medical applications. SemanticCT has been semantically integrated with various data, which include various trial documents with semantically annotated eligibility criteria and large amount of patient data with structured EHR and clinical medical records. Well-known medical terminologies and ontologies, such as SNOMED, LOINC, etc., have been used for the semantic interoperability.

SemanticCT is built on the top of LarKC (Large Knowledge Collider), a platform for scalable semantic data processing[2]. With the built-in reasoning support for large-scale RDF/OWL data of LarKC, SemanticCT is able to provide various reasoning and data processing services for clinical trials, which include faster identification of eligible patients for recruitment service and efficient identification of eligible trials for patients.

The contribution of this paper is: (1) a framework that enables semantic technologies for medical tasks related to the domain of clinical trials. (2) a proof of concept of the framework by SemanticCT with a focus on three tasks: (i) semantic search for clinical trials and patient data,(ii) trial finding for patients, (iii) identifying patients for a trial.

This paper is organized as follows: Section 2 presents the general ideas of semantically enabled systems for clinical trials. In section 3 we focus on three tasks in the clinical trial domain: search in clinical trials and patient data, trial finding for patients and identifying eligible patients for trials. Section 4 describes a formalization of eligibility criteria of clinical trials. Section 5 proposes the architecture of SemanticCT and describes various services and interfaces of the system. Section 6 discusses the related work and make the conclusions.

2 Approach

The goal of SemanticCT is to exploit semantic techniques in the domain of medical trials such that several tasks like trial finding and identifying eligible patients for trials can be supported. In this section we describe the semantic data integration and the platform. Notice that we use existing semantic technologies, available medical ontologies, data sources, and semantic annotaters.

2.1 Semantic Data Integration

Semantic data integration of various data in clinical trials is a basic step to build a semantically enabled system for clinical trials. Many existing trial data are usually represented as XML data with the standard fields. For example, the clinical trial service in the U.S. National Institutes of Health[3] provides the structured CDISC 20 fields of XML-encoded trial data. We can convert those XML data into standard semantic data, like RDF NTriple data with the annotations of medical ontologies or terminologies, like SNOMED, LOINC, MESH and others. Those ontologies can be used individually, or in a group with the

[2] http://www.larkc.eu
[3] http://www.clinicaltrials.gov

ontology alignments which are provided by the BioPortal ontology service[4] or other alignment tools. LinkedCT[5] provides large-scale semantic data of clinical trials with the standard formats of Linked Open Data in the Semantic Web.

The semantic annotations of clinical trials can be obtained by using many semantic annotation tools/systems, which have been developed by the community of the Semantic Web. BioPortal and MetaMap[6] provide satisfying services for semantic annotations with biomedical ontologies. Those annotation data are also represented as XML ones. Similarly it is easy to use XSLT to convert those XML encoded data into RDF NTriple ones. This means that for semantic interoperability we can exploit the available mappings among ontologies or to load an own allignment as RDF NTriples.

Some EHR prototype systems have been developed to support some kinds of semantics-enriched patient data. Those patient data can be accessed via the servers provided by those systems. However, real patient data are usually protected and not allowed for public access, because of the legal issue and privacy reason. We have developed a knowledge-based patient data generator which can synthesize required patient data for the purpose of tests by using some domain knowledge to control the data generation and make the generated data look like realistic ones[6].

In the paper we take three tasks into account: search in clinical trials or patient records, trial finding for patients, and identifying eligible patients for trials. For our feasibility tests we use clinical trials of breast cancer and we integrated the following data in SemanticCT:

Clinical Trials. We got the XML-encoded data of 4665 clinical trials of breast cancer from the official NCT website `www.clinicaltrials.gov`, and used XSLT to convert the XML-encoded data into RDF NTriple data, which consists of 1,200,565 triples and 335, 507 entities.

Medical ontologies. We got the latest release of SNOMED terminologies and converted them into RDF NTriple data. The concepts and definitions of converted SNOMED consists of 4,048,457 triples, which correspond with 2,046,810 entities.

Semantic annotations of clinical trials. We used the semantic annotation server of BioPortal to obtain the XML-encoded semantic annotations of the 4665 clinical trials with the medical terminologies/ontologies such as SNOMED, LOINC, HL7, MESH, RxNorm, and EVS. We converted the semantic annotation data into RDF NTriple. The total data size is about 3.0 GB. For the experiment, we load the semantic annotation data with the SNOMED concepts only. This part of data consists of 106,334 triples (454MB data).

Patient Data. We used APDG (Advanced Patient Data Generator), a knowledge-based patient data generator, to create 10,000 patient data of breast cancer, which cover the main properties of female breast cancer patients, like demographic data (e.g., gender and age), diagnosis, TNM stage (T for primary tumor, N for regional lymph nodes, and M for distant metastasis), hormone

[4] http://bioportal.bioontology.org/
[5] http://linkedct.org/
[6] http://metamap.nlm.nih.gov/

receptor status, e.g., the status of ER (Estrogen Receptor), PR (Progesterone Receptor), and HER2 (Human Epidermal Growth Factor Receptor 2), etc. We have collected the domain knowledge from medical literature (like PubMed) and web pages (like those from Wikipedia) and encoded those domain knowledge to control the generation of patient data and make the generated patient data look like realistic ones[6]. The generated patient data set consists of 660,000 triples.

Thus, the total loaded RDF NTriple data are over 6 million triples. It is sufficient for a demonstration prototype which runs at an ordinary laptop (dual core and 4GB memory) with extremely good performance. Most SPARQL queries in SemanticCT can be finished within one second. Thus, the time performance is not a big issue. What we concern mainly is whether such an approach can be used for supporting clinical trial tasks by developing a trial finding service and a patient recruitment service.

2.2 Semantic Platform

There have been several well-developed triple stores which can be used to serve as a semantic platform to build SPARQL endpoints for the services of querying over large-scale semantic data. Well-known triple stores are OWLIM[7] and Virtuoso[8]. Those triple stores usually support for basic RDFS reasoning over semantic data.

LarKC is a platform for scalable semantic data processing. OWLIM is used to be the basic data layer of LarKC. LarKC fulfills the needs in sectors that are dependent on massive heterogeneous information sources such as telecommunication services, biomedical research, and drug-discovery[4]. The platform has a pluggable architecture in which it is possible to exploit techniques and heuristics from diverse areas such as databases, machine learning, cognitive science, the Semantic Web, and others. LarKC provides a number of pluggable components: retrieval, abstraction, selection, reasoning and deciding. In LarKC, massive, distributed and necessarily incomplete reasoning is performed over web-scale knowledge sources[10]. One of our clinical trial task requires a new reasoning component (see section 4) which can be plugged in the LarKC platform.

3 Tasks in clinical trial domain

There are a large number of tasks in the domain of clinical trials. In this paper we focus on the tasks search, trial finding for patients and identifying eligible patients for trials with the main question in mind whether the approach of semantically enabled system for clinical trials can support those knowledge intensive tasks.

3.1 Search

SemanticCT provides various search services over large-scale integrated data: clinical trials, medical ontologies and patient data (see section 2.1). The semantic integration is realised by several available medical ontologies and mappings

[7] http://www.ontotext.com/owlim
[8] http://virtuoso.openlinksw.com/

between those ontologies from BioPortal. We also provide the service for browsing semantically annotated eligibility criteria of trials, search services for patient data browsing and specific patient finding, such as, show all triple-negative breast cancer patients. These search facilities are all realized by enabling semantic technologies into the domain of clinical trials.

3.2 Trial Finding for Patients

The trial finding service is one which searches for suitable trials for a given patient. Namely, based on the patient data, the system will check the requirement of clinical trials with the patient data to see whether or not the trial can be considered as a candidate trial for further deliberation by the patient and the clinician to make the decision. Some requirements (such as gender and age) have been structured in the original XML data. Some of those requirements are stated in the eligibility criteria (i.e., inclusion criteria or exclusion criteria), which are represented in natural language text. There are different approaches to deal with the information in text. We can either use SPARQL queries with regular expressions over eligibility criteria, or SPARQL queries directly over semantic annotations of eligibility criteria, or formalize the text by using some kind of formalization to make the structured eligibility criteria.

Given a patient data, it seems to be ideal to check if all the properties of a patient meets the requirements of a trial. However, we have found that it is not necessary, because checking with a few properties are sufficient to reduce significant amount of candidate trials and result in a small amount of trials for further deliberation.

For the experiment, we select just a small set of checking items, which consists of some structured fields, such as demographic data (gender and age), and some unstructured data (i.e., those in the text of eligibility criteria) such as stage, menopausal status, and hormone receptor status. The latter can be checked by using regular expressions with filters in SPARQL queries. Of course, we are interested in the trials which are currently recruiting, rather than those which have been completed. Thus, the initial SPARQL query of trial finding for a female patient aged 40 at stage 2 can be represented as follows:

```
PREFIX ...
select distinct ?ctid ?summary ?criteria
where {
?ct rdf:type sct:ClinicalTrial.
?ct sct:NCTID ?ctid.
?ct sct:EligibilityGender 'Female'.
?ct sct:OverallStatus "Recruiting".
?ct sct:EligibilityMinAge ?minage.
?ct sct:EligibilityMaxAge ?maxage.
?ct sct:BriefSummary ?summary.
?ct sct:EligibilityCriteriaTextblock ?criteria.
FILTER(?minage <= '40 Years'&& ?maxage >= '40 Years').
FILTER regex(str(?criteria), 'stage 2').}
```

In the query above, the regex 'stage 2' is used to match the stage in the eligibility criteria. The way of text matching is not sufficient to find all the targeted information. We can extend the regular expressions to cover various expressions which talk about the stage in natural language text. It is quite clear that we cannot exhaust all the expressions which talk about the stage in natural language text. Furthermore, the query cannot make a distinction between the text appears in inclusion criteria and that in exclusion criteria, unless we introduce more complex regular expressions which can detect the beginning and the ending of those criteria.

We add checking on more properties of patients, like menopausal status and hormone receptor status. That would reduce more candidate trials. Such reduction is very useful for clinicians. Table 1 summarizes the results of trial finding with those selected properties for 11 randomly selected tests. Actually each test represents a type of patients with their corresponding properties. From the table, we know that just a few property checking would reduce significant amount of candidate trials and result in only a few trials for further decision. The maximal number of candidate trials is 28 and the minimal number of candidate trials is 3. We have also detected the problem that some item checking by regular expressions cannot deal with negation information correctly, in particular, for those appear in exclusion criteria. For example, for checking on 'hormone receptor status', four trials have been mistakenly identified.

Patient ID	Age	Stage	Found Trial	Menopausal Status	RT	HR Status	RT	FF Trial	EF	Precision (%)
1000001	40	0	19	premeno	1	ER+,PR-,HER2+	2	16	1	93.75
1000302	67	2	16	postmeno	0	ER-,PR-,HER2+	2	14	0	100
1001422	61	1	11	perimeno	0	ER+, PR+, HER2-	0	11	0	100
1001548	64	1	11	postmeno	0	ER+, PR-, HER2-	0	11	0	100
1002017	52	2	18	perimeno	2	ER+,PR-,HER2+	1	15	0	100
1003862	69	0	32	postmeno	0	ER-,PR+,HER2+	4	28	0	100
1004121	42	1	17	perimeno	3	ER-,PR+, HER2-	4	10	0	100
1005035	41	0	19	premeno	1	ER-,PR+,HER2+	2	16	1	93.75
1006125	47	0	19	perimeno	1	ER-,PR-,HER2+	2	16	1	93.75
1007321	75	3	26	postmeno	0	ER-,PR-,HER2-	23	3	0	100
1009934	64	3	27	postmeno	0	ER+,PR-,HER2-	4	23	1	95.65
Average	56.55		19.55		0.73		4.4	14.82	0.36	97.90

Table 1. Trial Finding for Patient by SPARQL Queries with Regular Expressions. RT: Reduced Trials, HR: Hormone Receptor, FF Trials: Finally Found Trials, EF: Error Found

This feasibility test shows us that SPARQL queries with regular expressions are useful and promising to select trials for a specific patient.

3.3 Identifying Eligible Patients for Trials

Another task is to provide faster identification service of eligible patients for clinical trials. That requires the formalization of eligibility criteria, so that matching patient data with formalized eligibility criteria for automatic identification of clinical trials for patients. In [5] we propose a rule-based formalization for eligibility criteria, which is briefly discussed in the next section 4.

We have picked up 10 clinical trials randomly and formalized their eligibility criteria by using the rule-based formalization. We have tested the system for automatically identifying eligible patients for those selected trials. The system is able to find minimally 241 patients and maximally 750 patients out of the 10,000 patients for each trial, within less five seconds, for the system which is running on an ordinary laptop (dual core and 4GB memory)[5]. This formalization is also useful for trial finding service, because it can provide exactly matching on the data, without relying on exhaustive regular expression patterns. This feasibility test shows us that rule-based formalization of eligibility criteria for identifying eligible patients for trials is doable in an effective and efficient way. Clearly the next step is to set-up an experiment with real patient data and validation of the results with a clinician.

4 Rule-based Reasoning

For reasoning over various semantic data for clinical trials, SPARQL queries are not always powerful and flexible enough to specify complex requirements of eligibility criteria. In the experiments with automatic identification of eligible patients, we have observed that SPARQL queries with regular expressions are not always sufficient, for instance, for checking eligibility criteria.

For example, in order to check if an eligibility criteria require a patient of the stage 2 breast cancer, we have to use a regular expression to cover various expressions which talk about the stage in natural language text, like this:

```
FILTER regex(str(?criteria),
'stage 2|stage II |stage 0, 1, 2|stage I, II|stage IIa|stage IIb')
```

As we have discussed before, it is quite clear that we cannot exhaust all the expressions which talk about the stage in natural language text. Therefore, that would result in some eligibility criteria uncheckable at the run time (i.e., querying time). We have developed a rule-based formalization of eligibility criteria for clinical trials[5], so that eligibility criteria in natural language text can be processed offline, i.e., when their formalizations are generated.

Compared with existing formalizations, the rule-based formalization is more efficient and effective, because of the declarative form, easy maintenance, reusability and expressivity[5].

There exist various rule languages which can be used for the formalization of eligibility criteria. In the researches of artificial intelligence, logic programming

languages, like Prolog, are well known and popular rule-based languages. Several rule-based languages, like SWRL [9] and RIF[10], have been proposed for the semantics-enable rule-based language. In biomedical domain, the Arden syntax [11] has been developed to formalize rule-like medical knowledge. However, compared with logic programming language Prolog, both SWRL, RIF and the Arden syntax have very limited functionalities for data processing.

In SemanticCT, the rule-based formalization is developed based on the logic programming language Prolog. We select the SWI-Prolog[12] as the basic language for the rule-based formalization of eligibility criteria, because of its Semantic Web support and powerful processing facilities [8,9].

We formalize the knowledge rules of the specification of eligibility criteria of clinical trials with respect to the following different levels of knowledge: trial-specific knowledge, domain-specific knowledge, and common knowledge.

Trial-specific Knowledge Trial-specific knowledge are those rules which specify the concrete details of the eligibility criteria of a specific clinical trial. Those criteria are different from a trial to another trial.

Given a patient ID, we suppose that we can obtain its patient data through the common knowledge of the interface with SPARQL endpoints and its internal data storage. Thus, in order to check if a patient meets an inclusion criterion, we can check if its patient data meet the criterion.

Furthermore, we would not expect to check all the criteria with respect to the patient data, because some of those required data may be missing in the patient data. We introduce a special predicate getNotYetCheckedItems to collect those criteria which have not yet been formalized for the trial.

For example, the inclusion criteria in the trial NCT00002720 can be formalized as follows:

```
meetInclusionCriteria(_PatientID, PatientData, CT,
             NotYetCheckedItems):-
                     CT = 'nct00002720',
                     breast_cancer_stage(PatientData, '1'),
                     invasive_breast_cancer(PatientData),
                     er_positive(PatientData),
                     known_pr_status(PatientData),
                     age_between(PatientData, 65, 80),
                     postmenopausal(PatientData),
                     getNotYetCheckedItems(CT, NotYetCheckedItems).
```

Which states that the inclusion criteria include: i) Histologically proven stage I, invasive breast cancer, ii) Hormone receptor status: Estrogen receptor positive and Progesterone receptor positive or negative, iii) Age: 65 to 80, and iv) Menopausal status: Postmenopausal.

Domain-specific Knowledge Those trial-specific rules above may involve some knowledge which are domain relevant, i.e., the domain knowledge, which

[9] http://www.w3.org/Submission/SWRL/
[10] http://www.w3.org/TR/rif-overview/
[11] http://www.hl7.org/special/Committees/arden/index.cfm
[12] http://www.swi-prolog.org/

are trial independent. We formalize those part of knowledge which are relevant with domain knowledge in the libraries of domain-specific knowledge. For example, for clinical trials of breast cancer, we formalize the knowledge of breast cancer in the knowledge bases of breast cancer, a domain-specific library of rules.

An example of this type of knowledge is a patient of breast cancer is triple negative if the patient has estrogon receptor negative, progesterone receptor negative and protein HER2 negative status. It can be formalized in Prolog as follows:

```
triple_negative(Patient):- er_negative(Patient),
                           pr_negative(Patient),
                           her2_negative(Patient).
```

We consider patient data as a set of property-value pairs. A general format of patient data, called the PrologCMR format, is designed to be a list of property-value pairs. This general format of patient data is flexible to represent the data from different formats of CMRs, because we can design a CMR-specific interface to obtain the corresponding data via different data servers, which can be a SPARQL endpoint, internal data storage server, or a database server[5]. Then, we can convert the patient data into one in the PrologCMR format. We introduce the general predicate getItem(PatientData, Property, Value) to get the value of the property from the patient data.

For example, these receptor status can be straightforward formalized as follows:

```
er_positive(PatientData):- getItem(PatientData, er, ER),
                           ER = 'positive'.
```

Common Knowledge The specification of the eligibility criteria may involve some knowledge which are domain independent, like the knowledge for temporal reasoning and the knowledge for manipulating semantic data and interacting with data servers, e.g. how to obtain the data from SPARQL endpoints. We formalize the knowledge in several rule libraries, which can be reusable for different applications.

Example of this type of knowledge is temporal reasoning with constructs like last-month.

```
lastmonth(LastMonth):- today(Today),
                Today = date(_Year, ThisMonth, _Date),
                ThisMonth > 1,
                LastMonth  is ThisMonth - 1.

lastmonth(LastMonth):- today(Today),
                Today = date(_Year, ThisMonth, _Date),
                ThisMonth is 1,
                LastMonth  is 12.
```

Based on the SWI-Prolog's Web libraries, we can develop the interface with SPARQL endpoints to obtain semantic data (e.g. semantics-enable patient data and medical ontologies) for the rule-based formulation of eligibility criteria.

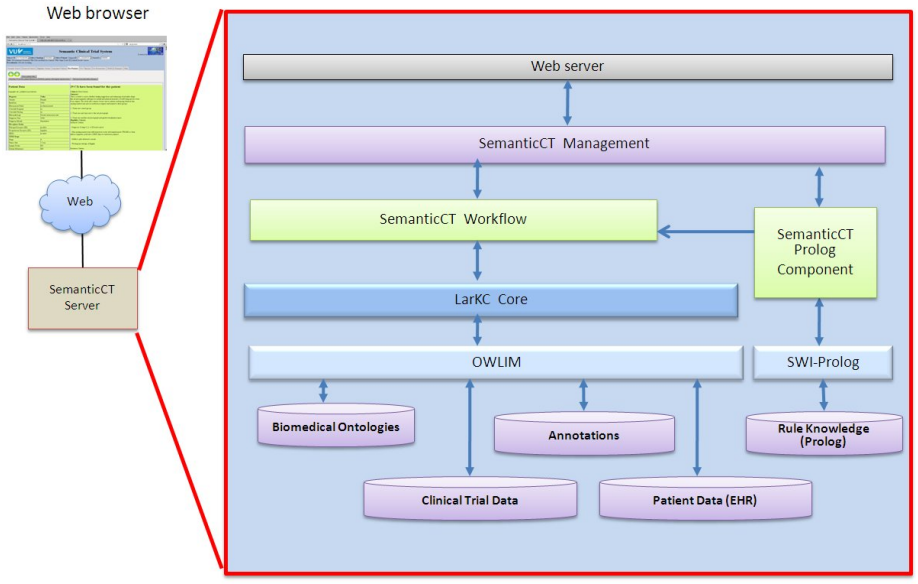

Fig. 1. The architecture of SemanticCT.

This reasoning component is developed as a LarKC component for the task of identifying eligible patients for trials. The rule-based reasoning component is also useful for trial finding service, because it can provide exactly matching on the data, without relying on exhaustive regular expression patterns. In the future we want to use this component for trial finding for patients. This requires that all eligibility criteria of the trials are modeled in this rule-based approach.

5 System

5.1 Architecture

The architecture of SemanticCT is shown in Figure 1. SemanticCT Management plays a central role of the system. It launches a web server which serves as the application interface of SemanticCT, so that the users can use a web browser to access the system locally (i.e., from the localhost) or remotely (i.e., via the Web). SemanticCT Management manages SPARQL endpoints which are built as SemanticCT workflows. A generic reasoning plug-in in LarKC provides the basic reasoning service over large-scale semantic data, like RDF/RDFS/OWL data. SemanticCT Management interacts with the SemanticCT Prolog component which provides the rule-based reasoning[5,3].

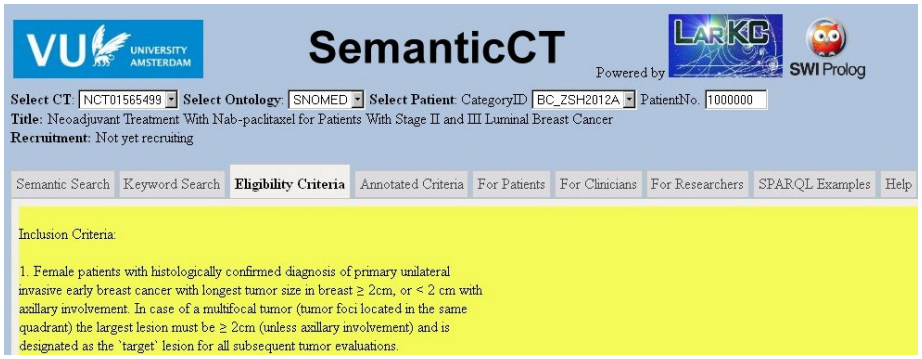

Fig. 2. The GUI of SemanticCT.

LarKC, which consists of the LarKC core for plug-in and workflow management and the LarKC data layer, serves as the infrastructure of SemanticCT for semantic data management. The LarKC data layer manages the semantic data repositories of SemanticCT. Those semantic data repositories consist of i) biomedical terminologies or ontologies, such as SNOMED CT, LOINC, MeSH, RxNorm, etc., ii) semantic data of clinical trials, like those from LinkedCT, or semantic data which are converted from the original XML-encoded data of clinical trials, iii) semantic annotation data of trials, which are generated from the biomedical semantic annotation servers, and iv) patient data, which can be the semantic data obtained from EHR systems, or created by the knowledge-based patient data generator[6]. Those semantic data repositories can be located locally or distributively.

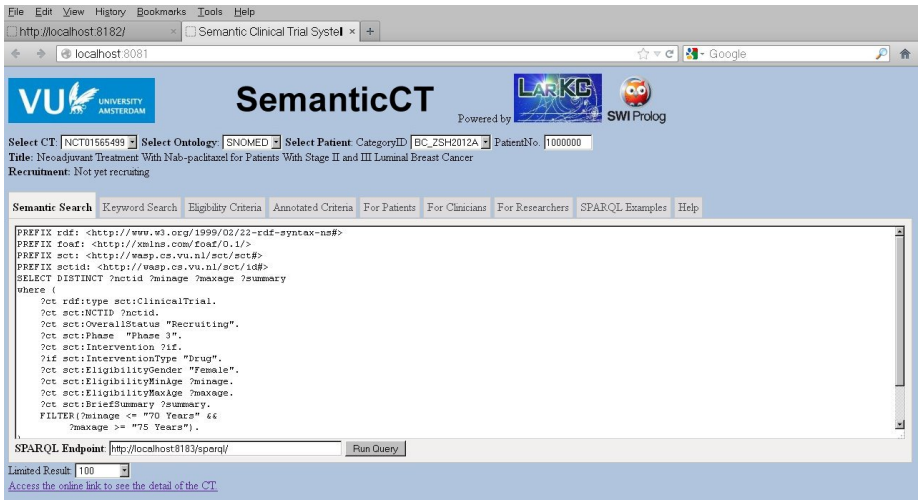

Fig. 3. The interface of semantic search.

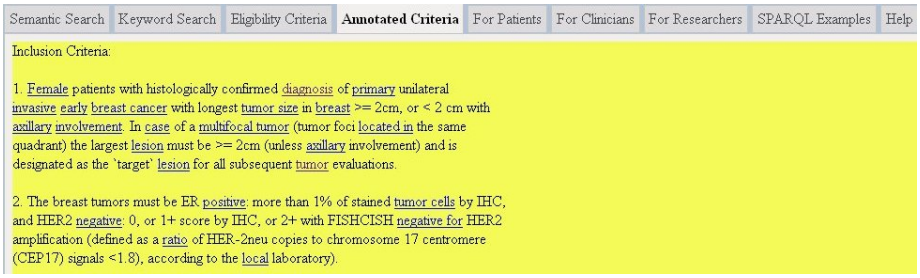

Fig. 4. Semantic Annotation

5.2 Interface and Service

For the demonstration prototype of SemanticCT, we merge the interfaces for various groups of users into a unique one on a Web browser, without considering their data protection issues, like access authority and password checking. A screenshot of the interface of the demonstration prototype SemanticCT is shown in Figure 2. Notice the several tabs that are available for various services and different types of users and discussed below.

- Semantic search: Figure 3 shows the interface of the semantic search, with a SPARQL example which searches for all recruiting phase 3 trials for female patients with the age between 70 and 75. We provide a set of SPARQL query templates, so that the users can select some of them and change some parameters of the templates to make their own queries (see SPARQL examples)
- Keyword search: We provide the ordinary search by using keywords to search over the eligibility criteria, or summaries of clinical trials. The extended keyword search provides complex keyword searches with the Boolean operators.
- Eligibility criteria: the eligibility criteria of the trial are shown.
- Annotated criteria: the service for browsing semantically annotated eligibility criteria of trials, see Figure 4.
- For patients: One of the main services in SemanticCT is the trial finding service. Currently, we provide the trial finding service by using SPARQL queries with regular expressions. The interface of patient services is shown in Figure 5. Notice that the SPARQL query is not visible for the user, but behind the button "show the CTs for this patient".
- For clinicians: SemanticCT provides several services for clinicians (see Figure 6). Those services include i) patient data browsing, ii) specific patient finding, such as, show all triple-negative breast cancer patients, and iii) patient recruitment for the selected clinical trial. The interface of clinician service for patient recruitment is shown in Figure 6. Notice that patient recruitment service is based on the rule-based formalization of the eligibility criteria.
- For Researchers: Semantic search for patient recruitment is one of the main services here.

Notice as well that the user can select an ontology from a list. In Figure 2 SNOMED is selected.

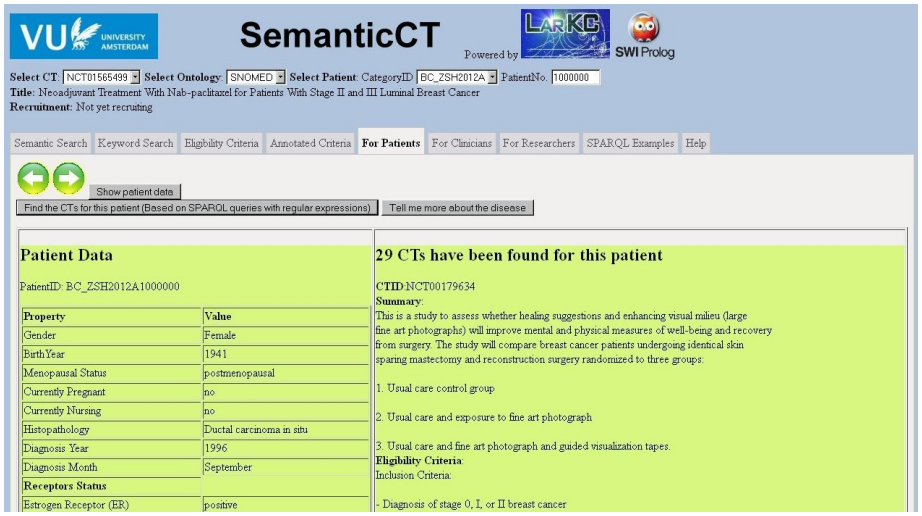

Fig. 5. Patient Service view and Trial Finding

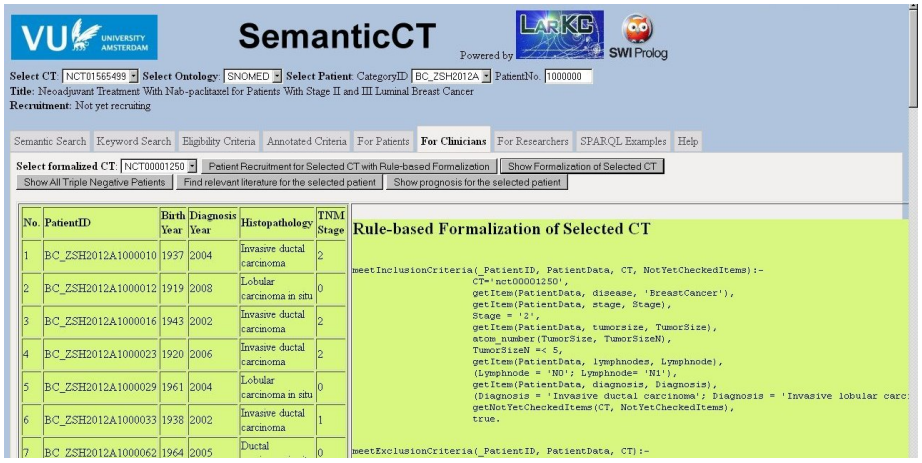

Fig. 6. Clinician services view and Rule-based formalization for Eligible Patient Identification

6 Discussion and Conclusion

6.1 Related Work

One of the obstacle to automate a clinical task like improving cohort selection for clinical trials is the need to bridge the semantic gap between raw patient data, such as laboratory tests or specific medications, and the way a clinician interprets this data. In [7] they presented a feasibility study for an ontology-based approach to match patient records to clinical trials. This is inline with

SemanticCT which enables to bridge this semantic gap as well by exploiting ontologies.

The work in [1] is also focused on the enabling of the semantic interoperability between clinical research and clinical practice. Their approach is based on a SOA-oriented approach combined with the exploitation of ontologies which forms an "intelligence" layer for interpreting and analyzing existing data, which is dispersed, heterogeneous information, which is to a great extend publicly available. In [2] the authors 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. They use a domain specific ontology to represent data from patients' health records and use SWRL to verify the eligibility of patients to clinical trials. Although we propose an even more expressive language (e.g., support for temporal reasoning and others) for modeling the eligibility criteria, this is in the same spirit as our approach. Furthermore, we use a general framework for specifying the eligibility criteria in three types of knowledge which can be reused.

6.2 Discussion

In this paper, we have presented a semantically-enabled system for clinical trials. We have proposed the architecture of SemanticCT, which have been designed to build on the top of LarKC, a platform for scalable semantic data processing. The logic programming language Prolog has been introduced to a rule-based formulization of eligibility criteria for clinical trials. SemanticCT has been semantically integrated with large-scale and heterogeneous data.

We have conducted several experiments for reasoning and data processing services over SemanticCT. The experiment of trial finding service shows that SPARQL queries with regular expressions are useful to deal with the information which can be easily obtained by the processing (like menopausal status and hormone receptor status). The experiment of the rule-based formalization shows that it is efficient and effective approach for faster identifying eligible patients. What we have implemented and tested is just a prototype of SemanticCT. Thus, it provides only a basic step for developing semantically enabled systems for clinical trials.

6.3 Future work

There are many interesting issues for future work of SemanticCT, which include trial finding by using rule-based reasoning, more comprehensive workflow processing for decision support procedure, deeper reasoning with biomedical ontologies, personalized information services for patients, clinicians, and researchers, etc. We are going to provide more extended services for clinicians, which include finding relevant and latest literature like those from PubMed for the selected patient, and showing prognosis for selected patients. The existing implemented prognosis service in SemanticCT is quite simple, for it shows only the 5 year survival rate, based on the TNM stage of patients. A comprehensive prognosis service would be able to make analysis of all the relevant patient data to finding

most-relevant clinical evidence for the prognosis analysis. We will continue the development of SemanticCT and deploy it in real application scenarios.

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. V. Andronikou, E. Karanastasis, E. Chondrogiannis, K. Tserpes, and T. A. Varvarigou. Semantically-enabled intelligent patient recruitment in clinical trials. In *Proceedings of International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 326–331, 2010.
2. P. Besana, M. Cuggia, O. Zekri, A. Bourde, and A. Burgun. Using semantic web technologies for clinical trial recruitment. In *International Semantic Web Conference*, pages 34–49, 2010.
3. A. Bucur, A. ten Teije, F. van Harmelen, G. Tagni, H. Kondylakis, J. van Leeuwen, K. D. Schepper, and Z. Huang. Formalization of eligibility conditions of CT and a patient recruitment method, D6.1. Technical report, EURECA Project, 2012.
4. D. Fensel, F. van Harmelen, B. Andersson, P. Brennan, H. Cunningham, E. Della Valle, F. Fischer, Z. Huang, A. Kiryakov, T. Lee, L. School, V. Tresp, S. Wesner, M. Witbrock, and N. Zhong. Towards LarKC: a platform for web-scale reasoning. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2008)*. IEEE Computer Society Press, CA, USA, 2008.
5. Z. Huang, A. den Teije, and F. van Harmelen. Rule-based formalization of eligibility criteria for clinical trials. In *Proceedings of the 14th Conference on Artificial Intelligence in Medicine(AIME 2013)*, 2013.
6. Z. Huang, F. van Harmelen, A. ten Teije, and K. Dentler. Knowledge-based patient data generation. In R. Lenz, S. Mikszh, M. Peleg, M. Reichert, and D. R. A. ten Teije, editors, *Process Support and Knowledge Representation in Health Care*. Springer LNAI, 2013.
7. C. Patel, J. J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg, and K. Srinivas. Matching patient records to clinical trials using ontologies. In *Proceedings of the International Semantic Web Conference*, pages 816–829, 2007.
8. J. Wielemaker, Z. Huang, and L. van der Meij. SWI-Prolog and the web. *Journal of Theory and Practice of Logic Programming*, (3):363–392, 2008.
9. J. Wielemaker, T. Schrijvers, M. Triska, and T. Lager. SWI-Prolog. *Journal of Theory and Practice of Logic Programming*, (1-2):67–96, 2012.
10. M. Witbrock, B. Fortuna, L. Bradesko, M. Kerrigan, B. Bishop, F. van Harmelen, A. ten Teije, E. Oren, V. Momtchev, A. Tenschert, A. Cheptsov, S. Roller, and G. Gallizo. D5.3.1 - requirements analysis and report on lessons learned during prototyping. Larkc project deliverable, June 2009.