Preprint
To appear IJSSCI 6(3), 2014

Patient-centered Clinical Trials Decision Support using Linked Open Data

Bonnie MacKellar
Division of Computer Science, Mathematics and Science
St John's University
Queens, NY
mackellb@stjohns.edu

Christina Schweikert
Division of Computer Science, Mathematics and Science
St John's University
Queens, NY
schweikc@stjohns.edu

Soon Ae Chun
Columbia University* &
CUNY College of Staten Island
New York, NY, USA
soon.chun@csi.cuny.edu
*On leave from CUNY College of Staten Island

***Abstract*—Patients often want to participate in relevant clinical trials for new or more effective alternative treatments. The clinical search system made available by the NIH is a step forward to support the patient's decision making, but, it is difficult to use and requires the patient to sift through lengthy text descriptions for relevant information. In addition, patients deciding whether to pursue a given trial often want more information, such as drug information. Our overall aim is to develop an intelligent patient-centered clinical trial decision support system. Our approach is to integrate Open Data sources related to clinical trials using the Semantic Web's Linked Data framework. The linked data representation, in terms of RDF triples, allows the development of a clinical trial knowledge base that includes entities from different open data sources and relationships among entities. We consider Open Data sources such as clinical trials provided by NIH as well as the drug side effects dataset SIDER. We use UMLS (Unified Medical Language System) to provide consistent semantics and ontological knowledge for clinical trial related entities and terms. Our semantic approach is a step toward a cognitive system that provides not only patient-centered integrated data search but also allows automated reasoning in search, analysis and decision making using the semantic relationships embedded in the Linked data. We present our integrated clinical trial knowledge base development and a prototype, patient-centered Clinical Trial Decision Support System that include capabilities of semantic search and query with reasoning ability, and semantic-link browsing where an exploration of one concept leads to other concepts easily via links which can provide visual search for the end users.**

Keywords—semantic web; Linked Open Data; clinical trials; knowledge representation

I. INTRODUCTION

Clinical trials are important, not just to researchers testing new treatments, but also to patients with serious diseases. In today's world in which patients take greater charge of their own health, it is common for patients to search the Internet for relevant clinical trials. In fact, the NIH makes all of its trials available on ClinicalTrials.gov for exactly this purpose. However, the task of sifting through lots of clinical trial descriptions to find the most appropriate one can be onerous [1]. There is a great need for developing a health knowledge repository that links different data from different sources scattered on the Web, with semantic relationships between data [36]. The semantic representation with explicit relationships can provide a system with intelligence, such as automated reasoning and search in supporting patient's decision on the clinical trials. The proposed system thus requires an interdisciplinary approach of Cognitive Informatics [35] that covers the areas of information modeling, of cognitive processes of human brain, and medical decision making.

Our overall aim is to build this kind of cognitive information processing system that allows patients to search open clinical trial information that returns only the trials that best fit a patient's characteristics and

information needs. In this paper, we focus on developing the clinical trial knowledge repository, which uses a Linked Open Data model to pull together data, both from the clinical trials and from other related datasets; in particular, side effect information. In this paper, we will give an overview of the architecture of this system, and then show how the ontology-based knowledge representation allows patients to view side effect information for treatments mentioned in particular clinical trials.

Patients and their caregivers search for clinical trials differently than physicians do. Patients typically do not have their own electronic health record (EHR), but often have a good recollection of their past treatments. If they moved between institutions during treatment, their knowledge may even be better than that in any one institution's records. The typical questions that patients may ask regarding the clinical trials include:

- What are the possible interventions that I might receive during the trial?
- What are the possible or known risks or side effects of this trial?
- Will the drugs I am taking have any negative interactions with one given in the trial?
- What tests and procedures are involved?
- Am I eligible to participate in this trial?

Patients' concern with side effects of treatments in a trial is shown in a very common query on patient-focused cancer discussion boards: "I am considering trial XYZ – can anyone let me know what the side effects are?" In order to answer some of these questions, the clinical trial data should be integrated with other data sources that represent drugs and their side effects or drug-drug interactions. Patients may also be intimidated by the need to read lengthy text-based descriptions of eligibility criteria and the need to type in search terms in text boxes [1]. Thus, a search tool aimed at patients should not only answer these patient-oriented questions, but rely on a more text-based interface without precise medical terms. By gaining insight into the cognitive process that a patient or caregiver goes through when searching for suitable clinical trials, we can better serve their needs. Rather than a patient having to repeat tedious steps on a variety of web resources, we can model this cognitive task in a comprehensive, connected, and guided system.

Our research develops a semantic web-based knowledge representation that integrates information about clinical trials from a number of Linked Open Data sources, in particular LinkedCT [19], a collection of clinical trials, and SIDER, a dataset that relates treatments to side effects. A semantic web knowledge representation is a powerful combination of web-based technologies with a graph-based knowledge representation known as the Resource Description Language or RDF [29]. This representation will be described more fully in Section III. Linked Open Data (LOD) is an approach to publishing, sharing, and linking datasets on the web using semantic web methods such as RDF and URIs, with the idea that information in these datasets can be easily queried and combined for new purposes [23]. These representations are augmented by linking medical concepts to UMLS (Unified Medical Language System) [2] concepts, to enable more powerful queries over the clinical trial information.

The paper is organized as follows. Section II surveys the related work on clinical trial search and current practices that identify the research issues involved. In Section III, we present the data sources we used for integrated search, and Section IV presents the data model of the Linked Data and the RDF representation of the integrated data sources. We present the prototype system architecture and components in Section V, followed by conclusions and future work in Section VI.

II. CURRENT APPROACHES TO CLINICAL TRIAL SEARCHES

Because the development of more effective search technology over clinical trials is so important, for both clinicians trying to accrue enough patients into a trial and for patients or their families who may be searching for the best clinical trial for their situation, there are quite a few researchers looking at the problem. For many years, clinical trial information in the United States was scattered across a number of websites, and made

available in HTML or PDF format, which is difficult to search. However, ClinicalTrials.gov was established by Congress in 1997 to make it easier for people to find trials, and by 2007, most non-phase 1 drug and interventional studies were included [3]. This site uses XML to represent the information in each trial description, with links to MESH terms, making automated searches far easier.

Most of the work centered on matching patients to trials has focused on the provider side of the equation, seeing the problem as one of automatically locating patients via electronic medical records that can be accrued onto a trial. The projects described by [4][12][14][15][16][15] are all examples. However, as noted above, patients and their caregivers also search for trials that match their own situation. Fewer researchers have focused on this scenario. Atkinson [1] reviewed a number of patient-oriented clinical trial search sites for usability, finding most of them difficult to use. An example of a system designed for patients is BreastCancerTrials.org [18], which was designed so that breast cancer patients could easily find clinical trials based on personal medical information entered by the patient. The OncoDoc system [6] mentioned earlier was also aimed at breast cancer patients and used a decision tree to guide the patients to selecting the best trial. Finally, the TrialX [19] site is also patient-focused, using patient data entered in Microsoft Health Vault to match against trials.

There are key differences between a system aimed at providers who need to accrue patients onto a trial, and a system aimed at patients who are searching for a relevant trial. Providers who are accruing patients are typically searching over the space of patient medical records within an institution. An example of this can be seen in [4] which describes a system that searches the electronic medical records of cancer patients at a cancer center for matches to a specific trial. Such systems need to automate the matching process since they are typically searching over a large number of patient records. A patient searching for trials, however, is usually searching based on his or her condition, which will typically result in a smaller number of results which can be then read by the patient. However, patients usually are interested in further information about the trial, such as locations, potential side effects (especially those which might limit participation in further trials), and whether the trial is currently accruing patients. A system which can integrate information on the trial from several sources, and which can highlight key information, allowing easy navigation, is more important to patients.

One problem is that significant parts of each trial description are still in free text; most notably the detailed description and the eligibility criteria. Furthermore, the language and formatting used in these sections is highly variable. Some studies divide eligibility criteria into separate inclusion and exclusion sections, or list each criterion on its own, bulleted line. Other trials simply list all criteria in one huge paragraph.

The eligibility criteria, however, are one of the most important factors when trying to decide if a particular trial is right for a patient. Typical criteria include factors such as patient age, pregnancy, tumor staging, existence of other conditions, and prior treatments. For example, trial NCT00945009, which is aimed at children with Wilms Tumor, stipulates that patients cannot have undergone nephrectomy at diagnosis, not be pregnant or nursing, and must have total bilirubin $\leq$ 1.5 times upper limit of normal for age. A clinician deciding which of his patients might be eligible, or a parent searching for the best treatment for his child would need to look at each criterion in turn to decide if the trial is a good fit or not. Thus, being able to search the eligibility criteria is a key feature and the focus of much research.

Systems aimed at providers need to automate matching of eligibility criteria to patient characteristics as much as possible since they must deal with a large number of patient records. Thus, many researchers are currently trying to solve this problem. One approach avoids the problem of free text by providing a structured language for trial designers to use when specifying a trial. Examples of this include a system for automating selection of patients for trials [4] which requires criteria to be entered in a logical language, a system that represents eligibility criteria as DL queries [5], and OncoDoc which represents eligibility criteria as nodes in a decision tree [6]. An example of a hybrid system is RuleEd [7] which is an editing environment in which clinicians enter eligibility criteria in free text. The text is parsed, and then the clinician works with this representation to

create a logical representation. Tu [8] reports on another system that semi-automates the translation of free text to a structured representation, in this case the ERGO language.

The second approach is to automate the process of translating the free text to a logical language. Researchers that are looking at the problem of completely automating the translation of eligibility criteria to a structured representation are mainly looking at small parts of the problem. For example, Luo et al [9][11] have investigated several components of the problem, including an algorithm for automatically inferring the semantic categories of eligibility criteria, the identification of Common Data Elements in eligibility criteria among multiple clinical trials studying the same disease, and the identification of temporal constraints. In another project, Lonsdale et al [12] used a set of clinical trials that had fairly structured free text criteria, extracting the criteria first into XML and then into a logic-based language. Milian, Bucur, and van Harmelen [13] present an approach which combines the idea of a structured language for specifying eligibility criteria at trial design time with the idea of mapping the free text to a logical language. In their approach, a library of eligibility criteria represented as an ontology is built by scanning the clinical trials in ClinicalTrials.gov and extracting the free text eligibility criteria.

Automated translation of free text in clinical trial descriptions, if ever solved, has the significant advantage of working with the large body of already written clinical trials, and does not require trial designers to learn a new way of entering trial specifications. However, this seems to be a problem that is not going to be solved soon. Since our system is oriented towards patients, fully automated matching of eligibility criteria to the patient record is not as critical. Therefore, our approach will be to simply identify the important keywords in the eligibility criteria, associating them with their UMLS concepts to make it easier for users to quickly identify trials that eliminate participants based on, for example, a particular prior treatment. We believe this will be enough to guide patients and prevent them from having to sift through lengthy eligibility criteria descriptions.

III. LINKED OPEN DATA

Linked Open Data utilizes RDF to describe data resources in a directed graph type of knowledge representation. This knowledge representation describes the link between information within a resource, as well as links to information located in external resources. This enables the cognitive properties of the information to be represented and queried. The SPARQL query language is designed to perform semantic queries on this knowledge representation by applying graph pattern matching between the query provided and the knowledge-based graph representation of the resource(s).

The registry of clinical trials housed at ClinicalTrials.gov [3], which is maintained by the National Library of Medicine (NLM) at the National Institutes of Health (NIH), serves as an online portal with information on clinical studies concerning a wide range of conditions that are conducted worldwide. Linked Clinical Trials (LinkedCT) is a semantic web resource of clinical trial data that can be semantically queried and provides links to external medical data sources **Error! Reference source not found.**. LinkedCT obtains its clinical trial dataset from ClinicalTrials.gov, and transforms this semi-structured data into RDF format. The clinical trial data in LinkedCT is continuously updated with new entries obtained from ClinicalTrials.gov. A valuable feature of LinkedCT is that it links conditions, interventions, trials, and references to related pages in other sources [20], such as DBpedia, Diseasome, DrugBank, DailyMed, ClinicalTrials.gov, and Bio2RDF's PubMed. Of course, since all of these data sources may have different identifiers and names for an entity, exact string matching will not be sufficient in locating all links to related information. To resolve this issue, LinkedCT utilizes approximate string matching and semantic matching to find related terms. Diseases and drugs may be referred to with different names across data sources. For example, the term "Heart Attack" may be used in one space, and "Myocardial Infarction" in another. In order to recognize that these two terms refer to the same concept, the NCI thesaurus is used to discover ontological relationships.

One of our goals is to facilitate a search on clinical trials that will be enhanced by including side effect information for drugs involved in the interventions of each clinical trial. Identifying potential side effects that may result from drug treatment is of great concern to potential clinical trial participants. We utilize a side

effect resource (SIDER) that contains drugs and their associated adverse drug reactions [21]. The data source was built from package inserts from multiple entities, including the US Food and Drug Administration (FDA). SIDER, in its current version, contained 996 drugs, 4192 side effects, and 99423 connections (http://sideeffects.embl.de/). The side effects come from a wide range of type, severity, and extent.

Applications and research in many fields, including biomedical and healthcare, can be greatly enhanced if data from various sources can be accessed and connected. Linked Open Data (LOD) is based on using the Semantic Web to create links between resources (structured or unstructured) that may be housed in any geographic location and maintained by different organizations [20][21]. Publishing and accessing Linked Data is done according to Linked Data principles, as defined by Berners-Lee [22], which involve using URIs (Uniform Resource Identifiers) for names, HTTP URIs to make these names accessible, providing relevant information using RDF and SPARQL standards, and including links to other URIs that contain useful information. RDF (Resource Description Framework) is used to describe and model the information of a web resource. The Linking Open Data (LOD) project [31] focuses on publishing publicly available data sets as linked data by converting them to RDF, utilizing the LOD standards, and connecting them to other data sets on the Web. There is an enormous amount of data in the life science and health domains that are housed in various organizations and in a variety of formats. Researchers can benefit from the ability to access, query, and make connections between these data sets, for example, relationships between genes, pathways and interactions, diseases, and drugs. The Linked Life Data platform interconnects over 20 data sources and enables users to perform semantically meaningful queries [23].

RDF [29] and SPARQL [30] are standards used throughout semantic web applications, including of course, Linked Open Data. The RDF standard represents information as a labeled directed graph. A concept, referred to as a resource, is linked to other concepts and properties via edges, leading to a representation based on triples: Subject-Predicate-Object, where each component is a resource identified by a Uniform Resource Identifier (URI) or a data literal such as a string. For example, the following two triples assert that there is a resource person#1233 named Bonnie MacKellar who is in the CUS department.

```
<ourdirectory:person#1233> <foaf:name> <”Bonnie MacKellar”>
<ourdirectory:person#1233> <ourdirectory:dept> <”CUS”>
```

The “foaf” alias refers to a URI for FOAF which is a popular ontology describing people and relationships [32]. SPARQL is a query language for data represented in RDF, which uses graph matching in order to retrieve results. A simple query might be

```
SELECT  ?d
WHERE ?p <foaf:name> “Bonnie MacKellar”.
      ?p <ourdirectory:dept> ?d.
```

This would return `d=”CUS”,` in other words, the department for Bonnie MacKellar. The query consists of the SELECT clause which specifies the results to return, and a WHERE clause which describes a graph pattern to match. The variable ?p will be bound to resources that match both triple patterns, and ?d will be bound to the subject in the triple that `matches ?p <ourdirectory:dept> ?d`.

In order to permit inference, two more standards have been developed that permit metadata to be represented: RDF Schema (RDFS) [33] and the Ontology Web Language (OWL) [34]. For example, RDFS adds resources for describing type (rdfs:type) and subclasses (rdfs:subClassOf), thus allowing type inference to occur. OWL is an extension of RDFS, permitting more complex type information to be specified, and is widely used to represent ontological information. Thus, we could assert that

```
<ourdirectory:person#1233> <rdfs:type> <ourdirectory:employee>
<ourdirectory:employee><rdfs:subClassOf> <foaf:person>
```

which states that the resource person#1233 is an employee, which is a subclass of “person” in the FOAF vocabulary.

Queries can proceed across distributed documents, much as searches on the current Internet proceed, leading to the ability to locate, retrieve, and integrate data which may exist on physically separate servers. This is at the core of the Linked Open Data paradigm. There are, however, a number of issues that can make data integration with LOD difficult. It is often difficult to understand a given dataset in the LOD cloud. The same concept may have different representations in different schemas and names may not match exactly. Data quality can be an issue since many of the datasets are converted into RDF using automated tools. Querying across multiple remote servers can be slow, and unfortunately there can be issues with downtime of endpoints. In order to minimize downtime and performance problems, we use a number of datasets from the Linked Life Data platform. We downloaded the needed datasets and store them locally in a Sesame triplestore.

A similar project [27] uses LOD as part of an RDF representation of electronic health record information, permitting querying of the patient data to identify cohorts with particular conditions. That particular project uses SIDER as well as some other LOD datasets, but does not work with clinical trials, nor is it integrated with UMLS, choosing instead to use the LOD Translational Medicine Ontology. Mucke et al [8] developed an RDF representation to model items in clinical trials, along with specifically developed OWL ontologies. This system, however, did not reuse linked open data or concepts in UMLS.

IV. KNOWLEDGE REPRESENTATION

A. *Linked Knowledge Representation Model*

In this section, we present our linked data model for the integrated clinical trial knowledge representation. This semantic representation of this linked data model serves as glue between the various LOD datasets. First, a conceptual view of the linked knowledge representation model is shown in Figure 1:

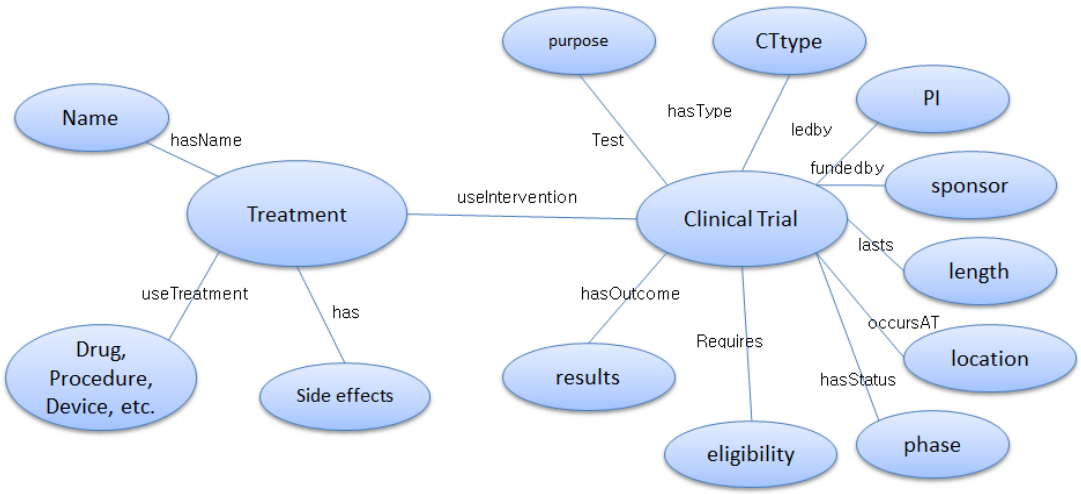

Figure 1: Linked KR Model for Integrated Clinical Trial knowledge base

By linking clinical trial information and side effect information, we will be able to generate queries over the knowledge base that answer many of the patient questions noted in the introduction.

In addition, the concepts or terms used in the linked data model are tied to UMLS concepts. The Unified Medical Language System (UMLS) provides a thesaurus of medical concepts as well as mappings to other controlled medical vocabularies. An implementation of the UMLS exists within the LOD cloud; however, the resources used in LinkedCT do not link to the UMLS concepts. The side effect resources in SIDER contain the concept identifiers (CUI) of the related UMLS concepts, but no direct RDF link to the UMLS concept itself. There are significant advantages to integrating our linked data concepts with the UMLS, such as consistent semantics among different data sources, and the type hierarchy-based reasoning capabilities.

For example, if a given trial specifies as part of its treatment protocol “monoclonal antibody 3F8”, integrating with UMLS gives us the knowledge that this concept has the type monoclonal antibody, and is a component of the chemotherapy regimen betaglucan/monoclonal antibody (C1134652), a concept whose semantic type is "Therapeutic or Preventive Procedure". This kind of reasoning capability will enhance the non-exact queries by patients that will yield appropriate query results, even though the exact keyword match does not exist. Therefore, our system adds explicit links between our resources and the associated UMLS concepts.

B. Linked Data Implementation with RDF Triples

Here we describe the RDF triple implementations of the linked data model using various data sources. The essential concepts in our linked data model are Treatment, SideEffect, and Clinical Trial. These tie together the concepts in SIDER, LinkedCT, and UMLS. A group of examples formatted using RDF/XML, follow in Figures 2, 3, and 5. This is for a clinical trial, NCT00697671, which is aimed at children with Chronic Myelogenous Leukemia, Acute Lymphoblastic Leukemia, Juvenile Myelomonocytic Leukemia, Myelodysplastic Syndrome, and Non-Hodgkin's Lymphoma (not all conditions are shown in the RDF below, for brevity). The first resource shown, in Figure 2, is a treatment, Etoposide, which is connected to the UMLS concept for Etoposide (C0015133). This treatment is used in the intervention protocol of the clinical trial NCT00697671, which is represented by the <usedInIntervention> link. There is a basic difference between the notion of an intervention and the notion of a drug. In LinkedCT, for example, there is a separate intervention resource for every treatment protocol that uses a drug. For example, “Etoposide, 100 mg/m^2 IV Daily Over 3 Hours x 4 Days” is an intervention in one clinical trial whereas “Etoposide, 40mg/m2, D1-4” which is part of a different clinical trial, is represented as a different intervention. In our representation, the Treatment resource represents one drug or procedure, rather than a protocol, and is also connected to possible side effects.

```
  <rdf:RDF
  <?xml version="1.0" encoding="UTF-8"?>
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
     <rdf:Description rdf:about="http:// trialbrowser/treatment1">
      <rdf:type rdf:resource="http://linkedlifedata.com/resource/umls/id/C0015133"/>
      <label xmlns="rdfs:">Treatment1 – Etoposide </label>
      <name xmlns="http://trialbrowser:">Etoposide</name>
      <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#"
rdf:resource="http://www4.wiwiss.fu-
      berlin.de/sider/resource/drugs/3310"/>
      <usedInIntervention xmlns="http://trialbrowser/"

  rdf:resource="http://data.linkedct.org/resource/intervention/23287"/>
      <hasSideEffect xmlns="http://trialbrowser/"

  rdf:resource="http://trialbrowser/sideEffect/sideEffect333"/>
  </rdf:Description>
```

Figure 2: Example of Treatment1 - Etoposide

```
  <rdf:Description rdf:about="http://trialbrowser/sideEffect/sideEffect333">
      <rdf:type rdf:resource="http://linkedlifedata.com/resource/umls/id/C0036572"/>
      <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#"
rdf:resource="http://www4.wiwiss.fu-

berlin.de/sider/resource/side_effects/C0036572"/>
  </rdf:Description>
```

Figure 3: Example of Side Effect triple in RDF

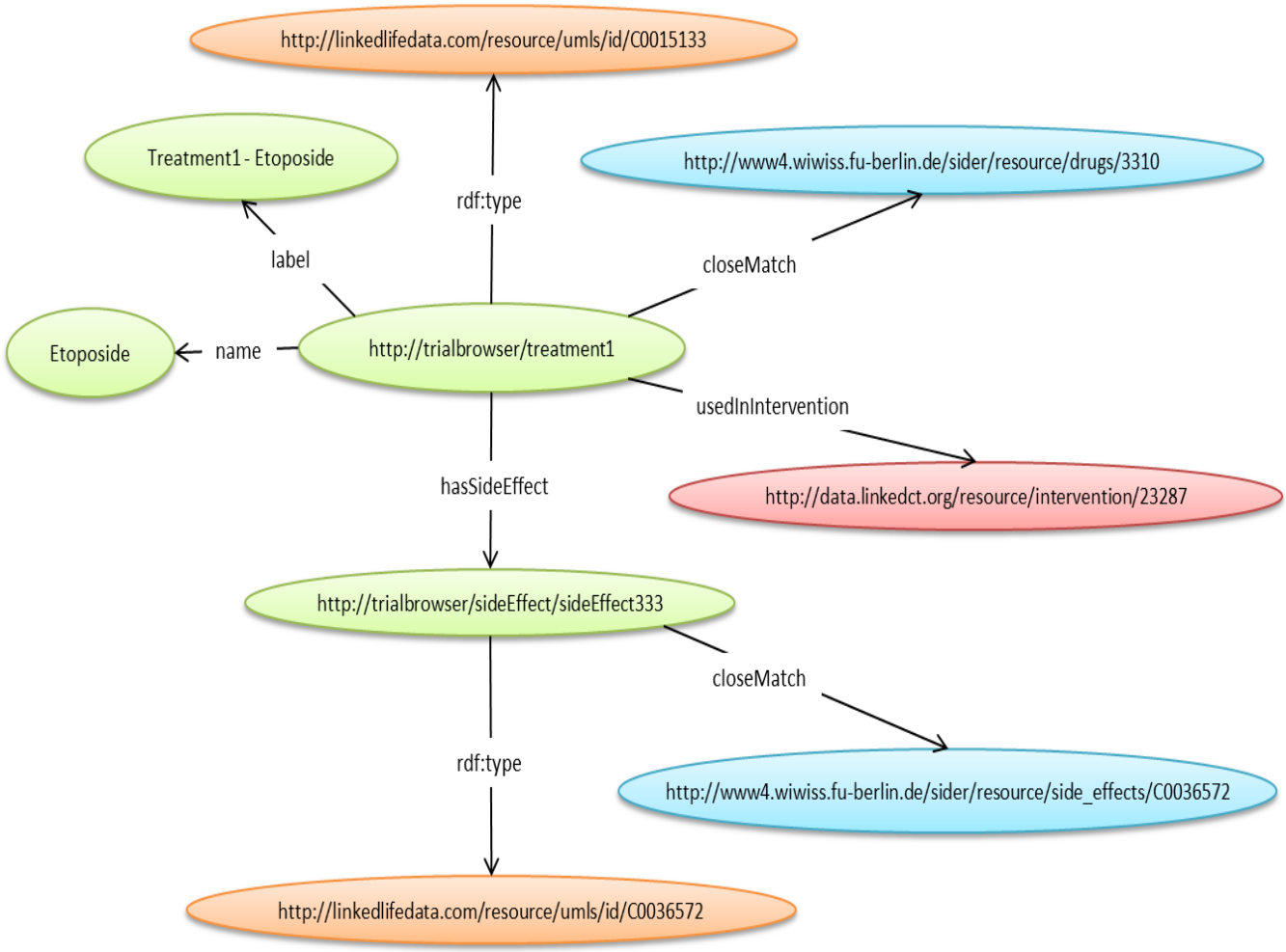

Figure 4: RDF triple representation of Treatment Etoposide and the side effect

The second resource, shown in Figure 3, is a side effect of etoposide, seizures, which is connected to the UMLS concept for seizures that has C0036572 as its concept id (http://linkedlifedata.com/resource/umls/id/C0036572) as well as to the SIDER side effect resource (http://www4.wiwiss.fu-berlin.de/sider/resource/side_effects/C0036572). Note that although the SIDER resource uses the UMLS Concept ID (C0036572), it is not the same resource, and has a different URI. All side effects of a treatment that are present in SIDER are included. Figure 4 shows an easier-to-read graphical representation of the RDF triple of Treatment1 linked with the SideEffect333.

The last resource, in Figure 5, is the clinical trial NCT0069767. This contains a link to the UMLS type Clinical Trial (C0008976), to the LinkedCT resource for the clinical trial, to the UMLS concepts for the conditions being treated in the trial, and for UMLS concepts listed in the eligibility criteria. This particular trial mentions pericardial effusion, radiation therapy and chemotherapy in the eligibility criteria. After tagging with MetaMap, we obtain the UMLS concept ids, which are used to generate the links shown below. Although far more information is associated with a clinical trial, our representation does not need to include this information because it is associated with the LinkedCT resource which is far more expansive. Figure 6 illustrates the interlinking of the clinical trial resource and its relationship with the three Linked Data datasets LinkedCT, SIDER, and UMLS.

```
<rdf:Description rdf:about="http://trialsbrowser/trialNCT00697671">
```

```
      <label xmlns="rdfs:">TrialNCT00697671</label>
      <rdf:type rdf:resource="http://linkedlifedata.com/resource/umls/id/C0008976"/>
      <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#"

rdf:resource="http://data.linkedct.org/resource/trials/NCT00697671"/>
      <trialCondition xmlns="http://trialbrowser/"

rdf:resource="http://linkedlifedata.com/resource/umls/id/C1327920"/>
      <trialCondition xmlns="http://trialbrowser/"

rdf:resource="http://linkedlifedata.com/resource/umls/id/C0751606"/>
      <eligibilityCriteriaContains xmlns="http://trialbrowser/"

rdf:resource="http://linkedlifedata.com/resource/umls/id/C0807677"/>
      <eligibilityCriteriaContains xmlns="http://trialbrowser/"

rdf:resource="http://linkedlifedata.com/resource/umls/id/C0805598"/>
      <eligibilityCriteriaContains xmlns="http://trialbrowser/"

rdf:resource="http://linkedlifedata.com/resource/umls/id/C0031039"/>
  </rdf:Description>
```

Figure 5: Example of Clinical Trial triple in RDF

Figure 6: RDF triple representation of Clinical Trial

In order to generate the RDF triples in the knowledge representation, two main methods are used. Most of the data can be retrieved via SPARQL queries. Whenever possible, direct links are used to tie together data, but this is not always possible. SIDER has the UMLS concept IDs associated with side effect resources, so these can be used to locate the actual UMLS resource and link it to our representation of the side effect. However, SIDER does not list the UMLS concept ID for drugs, and LinkedCT does not use them at all. This can be unreliable of course. The UMLS resource lists the alternative names used in the UMLS vocabularies. For example, Etoposide is also known as Demethyl Epipodophyllotox in Ethylidine Glucoside in MESH, and as EPEG in the NCI Thesaurus. Therefore, we can search all of the alternative names in order to determine which UMLS concept is the matching one. If a match still cannot be made, the Metamap concept tagger [28], which can identify matches to UMLS concepts in free text, can be used to locate the UMLS concept.

V. SYSTEM OVERVIEW

A. *Architecture of proposed system.*

This section describes an overall architecture and the major components of the proposed system shown in Figure 7. The prototype is implemented using a Sesame triple store and will have a Web-based interface for the patients.

The RDF triple generator uses an extractor component to generate the integrated linked data for representing the clinical trials knowledge base for patient oriented search. The triples are stored in the Sesame triplestore. The triples generated also use the external medical ontology, i.e. UMLS, to link the extracted data with the medical concepts and type information.

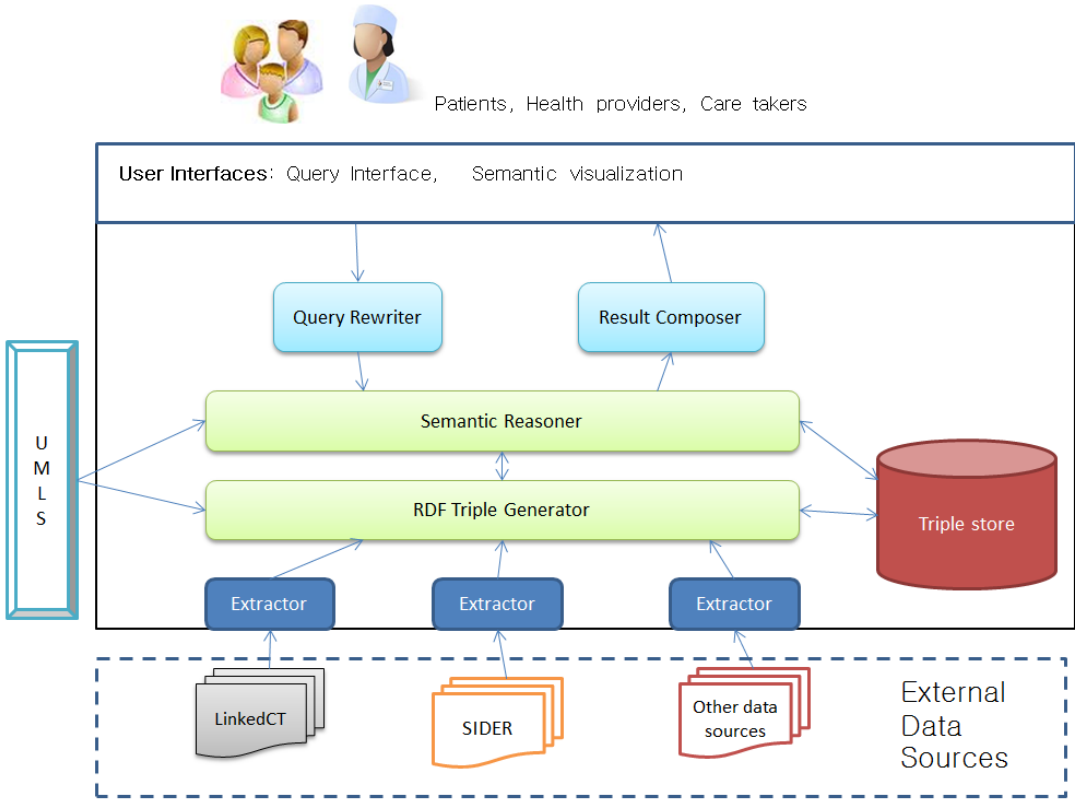

Figure 7: Prototype System Architecture for Integrated Clinical Trial linked with other data sources

One of its chief tasks of the RDF triple knowledge representation generator is to discover related concepts in the linked open data sets. These datasets are updated on a periodic basis, so our knowledge representation

must evolve as well. The triple generator links side effects of drugs and intervention for a certain condition. For performance reasons, we use dumps of these datasets loaded into our own triplestore, which are updated as new dumps become available. The extractors used by the knowledge representation generator are a set of SPARQL queries to extract the needed information. For example, to generate the side effects for a drug-related intervention for the condition "Chronic Myelogous Leukemia," the following extraction query can be used:

```
SELECT ?sideEffect ?intervention
  WHERE {
     ?trial linkedct:intervention ?intervention .
    ?intervention linkedct:intervention_name ?interventionname .
      ?trial linkedct:condition ?condition.
    ?condition linkedct:condition_name "Chronic Myelogous Leukemia" .
    ?drug sider:drugName ?interventionname .
            ?drug sider:sideEffect ?sideeffect .
```

Figure 8: Sample SPARQL query to obtain side effect information for a treatment

This works based on string matching between the SIDER representation for a drug name and the LinkedCT representation for an intervention name. This matching can be improved by using the UMLS concept information. However, neither SIDER nor LinkedCT link drugs or treatments to their UMLS concept. To do this, we utilize another SPARQL query which matches the treatment name against all alternate names recorded in UMLS.

The other major task of the knowledge representation generator is to index the eligibility criteria in each LinkedCT clinical trial using MetaMap. The results of this process will be used to generate the `<eligiblityCriteriaContains>` links in the knowledge representation.

The user interface and search component will consist of these subcomponents:

- A user interface and visual data explorer. An end user query may be in simple keywords, a form-based or visual navigational interface.
- A query generator which will convert input from the user interface to SPARQL queries.
- The Semantic Reasoner component helps to expand the query with necessary type or additional concepts using UMLS before it is sent to retrieve the relevant information.
- A results composer that converts query results into a user-friendly format, including visual display. For example, if a user, after viewing the results of a search, decides to click on the tag for all trials that have an eligibility criterion mentioning radiation, the query generator will generate the SPARQL query searching for all trials with a `<eligibilityCriteriaContains> radiation` link.

B. Use case analysis

In this subsection, we discuss use cases of our system.

a) Semantic search for clinical trials information

Potential clinical trial participants or their family members will be able to query the system and search for trials according to various terms or combinations of terms, including medical condition, intervention, drug, etc. For example, a user might search for side effects related to clinical trials related to "Chronic Myelogenous Leukemia." The user query can be mapped to UMLS concepts in order to broaden the search to include synonymous or similar terms by the Semantic Reasoner, in case the exact match does not work. A list of trials containing these concepts will be retrieved.

Figure 9 illustrates how the clinical trial and drug information fits into the UMLS Semantic Network. In this example, we are focusing on the conditions and interventions of clinical trials, along with the drugs and side effects in the side effect database. Our system's capability to use semantic reasoning on clinical trial and side effect information is enabled by incorporating knowledge from the UMLS Semantic Network, such as semantic types and relations. A clinical trial's condition(s) can be mapped to the semantic type Disease or Syndrome, or one of its subtypes. For example, the condition "Chronic Myelogenous Leukemia" and other Leukemia will map to the semantic type Neoplastic Process, which is a subtype of Disease or Syndrome. Some conditions, such as Crohn's disease map directly to the Disease or Syndrome type, and others may map to other subtypes of Disease or Syndrome, such as Dementia which maps to the disease subtype Mental or Behavioral Dysfunction. Interventions contain specific treatments that involve a Therapeutic or Preventative Procedure that uses a Pharmacologic Substance. Thus, when the user query is underspecified, as in this query "what are the pharmacologic substances that treat condition X?" where a super class concept (i.e. Pharmacologic Substance) is used in the query, the system can perform the reasoning and return the drug information (i.e. Pharmacologic Substances). The concept Etoposide, mentioned earlier, maps to the type Pharmacologic Substance. Side Effects can be of several different semantic types, such as Sign or Symptom, Finding, Disease or Syndrome, Pathologic Function, among others. The SIDER database reports chills as one of the side effects for the drug Etoposide; chills maps to the semantic type Sign or Symptom, which is a subtype of Finding. Another side effect, Alopecia, maps to Finding, and the side effect Leukopenia maps to the type Disease or Syndrome. The relationships between the information contained within clinical trials and side effect resources is complex and the UMLS semantic network helps to navigate these connections and make semantically rich queries.

Pulling all of this information into a semantic framework enables the system to provide the patient with a cohesive, integrated, and easy to navigate presentation. This will ease the search process for clinical trials in the number of searches that need to be conducted and the amount of free text that would have to be read. When searching for a disease on the ClinicalTrials.gov site, users are presented with a lengthy list of trials, some of which have already been completed or terminated. The records also contain detailed medical terminology that may be hard for the user to parse. If a user wishes to research side effects of a drug, he or she would need to conduct an additional search on the Internet for side effect information, which may be time consuming and difficult to find. In addition to semantic browsing, the system we propose will contain a guided series of questions that will lead to a smaller subset of clinical trials to consider. The semantic presentation will also greatly enhance the user's experience by providing more linked information and a more visual and user-friendly search.

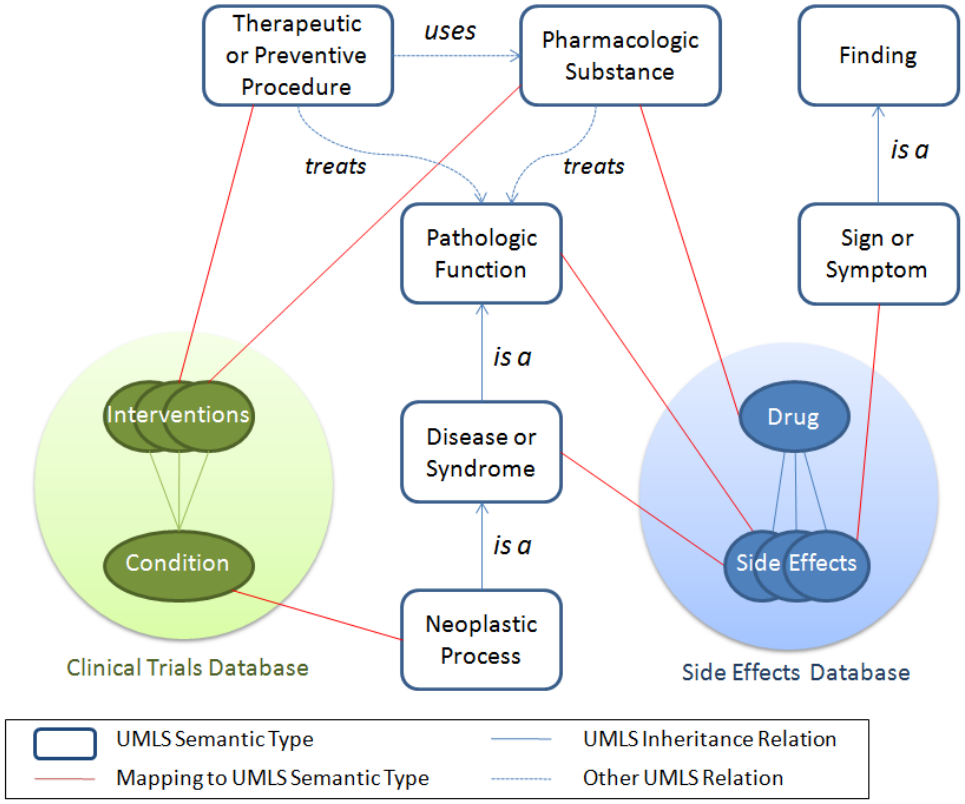

Figure 9: Clinical trials, drugs, and side effects from the perspective of the Semantic Network

b) *Semantic Link Browsing:*

Patients want to find out all clinical trials containing a certain drug in the eligibility criteria. It is difficult to navigate all the clinical trials involving the drug to identify eligibility criteria. Our system can allow browsing of concepts to navigate to other related concepts, which we call "hyper-semantic browsing." The contents of the clinical trial eligibility criteria, which are mostly textual, can be mapped to UMLS concepts using MetaMap. These concepts can be used as tags to build a tag cloud of concepts mentioned in the eligibility criteria. Thus, a click on the tag for drug Cisplatin will get all trials that mention the drug in the eligibility criteria.

Since we are using UMLS concepts, we can exploit the concept relations to create sub-groups within the eligibility criteria. For example, conditions and drugs can be grouped according to type. This enables a user to click on a concept group, such as "All platinum based drugs" and then trials that mention Carboplatin as well as Cisplatin would be retrieved. When a specific clinical trial or intervention is selected, the user will be presented with the treatment, related side effects, along with links to more information. By providing this side effect information to the user, directly alongside the clinical trial, this system would expedite the process of comparing clinical trials and facilitate an informative analysis that takes drug side effects into consideration.

VI. CONCLUSION

We have presented a semantic integration approach that allows patient-oriented search for information on clinical trials. The integration is based on LOD principles and semantic technologies, and is represented via

RDF triples in an integrated knowledge representation. We leveraged LOD sources such as the LinkedCT clinical trials dataset, the SIDER drugs and side effects dataset, and the UMLS medical ontology for consistent semantics across concepts used in different data sources. Our prototype system architecture includes a *knowledge generation* component where RDF triples are generated by extracting data from different sources and linking them with semantic information; a user interface that provides patients with powerful *semantic search* capabilities that use query processor and semantic reasoning components; and *semantic-link browsing* (a la hyperlink browsing) where the navigation from one concept to another can help users with visual search and exploration of clinical trials and related information. We presented use cases to illustrate the system functions such as query processing steps, semantic-search, and semantic-link browsing.

Our future research plan includes expanding the linked data to other sources, and enhancing the implementation of the prototype system. A user evaluation study of the system will be conducted. Additional features will be added to the system, such as helping patients determine which clinical trials they would be excluded from if they participate in a particular trial and, if they are eligible for multiple trials, which trial would be the best starting point. This patient-centered decision support system could also be used to monitor the way patients navigate through the data in order learn what is most important to them in the cognitive process of searching for clinical trials.

References

[1] N. L. Atkinson, S. L. Saperstein, H. a Massett, C. R. Leonard, L. Grama, and R. Manrow, "Using the Internet to search for cancer clinical trials: a comparative audit of clinical trial search tools.," *Contemporary Clinical Trials*, vol. 29, no. 4, pp. 555–64, Jul. 2008.

[2] K. E. Campbell, D. E. Oliver, and E. H. Shortliffe, "The Unified Medical Language System," *Journal of the American Medical Informatics Association*, vol. 5, no. 1, pp. 12–16, 1998.

[3] R. M. Califf, D. A. Zarin, J. M. Kramer, R. E. Sherman, L. H. Aberle, and A. Tasneem, "Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010.," *JAMA : the Journal of the American Medical Association*, vol. 307, no. 17, pp. 1838–47, May 2012.

[4] E. Fink, P. K. Kokku, S. Nikiforou, L. O. Hall, D. B. Goldgof, and J. P. Krischer, "Selection of patients for clinical trials: an interactive web-based system.," *Artificial Intelligence in Medicine*, vol. 31, no. 3, pp. 241–54, Jul. 2004.

[5] C. Patel et al., "Matching Patient Records to Clinical Trials Using Ontologies," IBM Research Report, 2007.

[6] B. Séroussi and J. Bouaud, "Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials," *Artificial Intelligence in Medicine*, vol. 29, no. 1–2, pp. 153–167, Sep. 2003.

[7] B. Olasov and I. Sim, "RuleEd, a web-based semantic network interface for constructing and revising computable eligibility rules.," *Proceedings of the Annual AMIA Symposium*, p. 1051, Jan. 2006.

[8] S. W. Tu et al., "A practical method for transforming free-text eligibility criteria into computable criteria.," *Journal of Biomedical Informatics*, vol. 44, no. 2, pp. 239–50, Apr. 2011.

[9] R. Mucke, M. Lobe, M. Knuth, et al, "A semantic model for representing items in clinical trials," *22nd IEEE International Symposium on Computer-Based Medical Systems*, Albuquerque, NM, USA: 2009, pp. 1–8.

[10] Z. Luo, S. B. Johnson, A. M. Lai, and C. Weng, "Extracting Temporal Constraints from Clinical Research Eligibility Criteria Using Conditional Random Fields" in *Proceedings of the Annual AMIA Symposium*, 2011, pp. 843–852.

[11] Z. Luo, R. Miotto, and C. Weng, "A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria.," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 33–39, Jul. 2012.

[12] D. W. Lonsdale, C. Tustison, C. G. Parker, and D. W. Embley, "Assessing clinical trial eligibility with logic expression queries," *Data & Knowledge Engineering*, vol. 66, no. 1, pp. 3–17, Jul. 2008.

[13] K. Milian, A. Bucur, and F. Van Harmelen, "Building a library of eligibility criteria to support design of clinical trials," *Knowledge Engineering and Knowledge Management, Lecture Notes in Computer Science*, vol. 7603, pp. 327–336, 2012.

[14] V. Andronikou, E. Karanastasis, E. Chondrogiannis, K. Tserpes, and T. Varvarigou, "Semantically-enabled Intelligent Patient Recruitment in Clinical Trials," *2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pp. 326–331, Nov. 2010.

[15] P. Besana, M. Cuggia, O. Zekri, and A. Bourde, "Using semantic web technologies for clinical trial recruitment," in *9th International Semantic Web Conference (ISWC)*, 2010, no. December, pp. 39–49.

[16] M. Cuggia, P. Besana, and D. Glasspool, "Comparing semi-automatic systems for recruitment of patients to clinical trials.," *International Journal of Medical Informatics*, vol. 80, no. 6, pp. 371–88, Jun. 2011.

[17] Y. Lee, D. Dinakarpandian, N. Katakam, and D. Owens, "MindTrial: An Intelligent System for Clinical Trials," *Proceedings of the Annual AMIA Symposium*, vol. 2010, pp. 442–6, Jan. 2010.

[18] E. Cohen et al., "Adoption, acceptability, and accuracy of an online clinical trial matching website for breast cancer.," *Journal of Medical Internet Research*, vol. 14, no. 4, p. e97, Jan. 2012.

[19] C. Patel, K. Gomadam, S. Khan, and V. Garg, "TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, no. 4, pp. 342–347, Nov. 2010.

[20] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R.J. Miller, M. Wang, "LinkedCT: A Linked Data Space for Clinical Trials," CoRR abs/0908.0567, 2009

[21] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs.," *Molecular Systems Biology*, vol. 6, p. 343, Jan. 2010.

[22] C. Bizer, T. Heath, T. Berners-Lee, "Linked Data—The Story So Far," *International Journal on Semantic Web and Information Systems* 5 (3): 1–22, 2009.

[23] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space," *Synthesis Lectures on the Semantic Web: Theory and Technology*, February 2011, vol. 1, no. 1 , pp. 1-136.

[24] T. Berners-Lee, "Linked Data - Design Issues," 2006, Retrieved March 1, 2013, http://www.w3.org/DesignIssues/LinkedData.html

[25] V. Momtchev, D. Peychev, T. Primov, G. Georgiev, "Expanding the Pathway and Interaction Knowledge in Linked Life Data," In *Proceedings of International Semantic Web Challenge*, 2009.

[26] M. Dumontier, "Building an effective Semantic Web for health care and the life sciences," *Semantic Web*, vol. 1, pp. 131–135, 2010.

[27] J. Pathak, R. C. Kiefer, and C. G. Chute, "Using semantic web technologies for cohort identification from electronic health records for clinical research.," *Proceedings AMIA Summit on Translational Science*, vol. 2012, pp. 10–9, Jan. 2012.

[28] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 17, no. 3, pp. 229–36, 2010.

[29] World Wide Web Consortium (W3C), "RDF - Semantic Web Standards." [Online]. Available: http://www.w3.org/RDF/. [Accessed: 06-Aug-2013].

[30] World Wide Web Consortium (W3C), "SPARQL Query Language for RDF." [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/. [Accessed: 30-Aug-2013].

[31] World Wide Web (W3C) Consortium, *The Linking Open Data Project*, retrieved March 17, 2013, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[32] "The Friend of a Friend (FOAF) project | FOAF project." [Online]. Available: http://www.foaf-project.org/. [Accessed: 30-Aug-2013].

[33] World Wide Web Consortium (W3C), "RDF Vocabulary Description Language 1.0: RDF Schema" [Online]. Available: http://www.w3.org/TR/rdf-schema/. [Accessed: 30-Aug-2013].

[34] World Wide Web Consortium (W3C), "Web Ontology Language (OWL)" [Online]. Available: http://www.w3.org/2001/sw/wiki/OWL. [Accessed: 30-Aug-2013].

[35] Y. Wang, et al, Perspectives on Cognitive Informatics and Cognitive Computing. Int. J. Cogn. Inform. Nat. Intell. 4, 1 (January 2010), 1-29, 2010.

[36] S. A. Chun and B. MacKellar, "Social health data integration using semantic Web," *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012 pp 392-397.