Semantic Integration of Software Systems in Translational Clinical Trials

Ravi D. Shankar[1], Martin J. 0'Connor[1], David B. Parrish[2], Amar K. Das[1]

[1]Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine
[2]The Immune Tolerance Network, Pittsburgh, PA
[1]{rshankar, martin.oconnor, das}@stanford.edu
[2]dparrish@immunetolerance.org

Informatics of Translational Clinical Trials

The translational research enterprise requires bi-directional sharing of data, knowledge, and information between researchers in the biosciences and those in clinical disciplines. Informatics efforts in translational research have focused largely on developing automated methods to correlate the results of genomics, proteomics, or mechanistic assay studies with available data on diagnosis, treatment, and outcomes. Often, the latter set of measures involves crude categories, such as 'cancer' or 'no cancer,' because further details about a patient's health or performed interventions are lacking. Such limitations create problems of specificity for findings in translational research. There are recent efforts to gather clinical information through standardized Electronic Medical Record representations that include genetics and genomics data [Hoffman 2007]. Such passive observations may yield biological insights into the mechanism of human disease and therapeutics. However, formalized controlled experiments, particularly human clinical trials, are necessary to address potential biases in biomarker analysis [Ransohoff 2005]. As a result, researchers are proposing and undertaking trial designs that include adjunct high-throughput assays or that directly evaluate biological hypotheses. We refer here to such studies as *translational clinical trials*.

The successful undertaking of any type of clinical trial requires significant work in knowledge specification, information management, and organizational coordination from the planning stage to final analysis. For more than a decade, informatics researchers have pursued a number of modeling projects that target system design requirements in trial registry, trial authoring, and trial execution. There are also major efforts supported by HL7[1] and CDISC[2] standards committees, in partnership with entities such as National Cancer Institute's caBIG project, to develop the BRIDG model[3] [Weng et al. 2007], which defines software functions and behaviors in a range of clinical trial applications, such as scheduling visits for a patient study calendar. This work is aimed at harmonizing standards within clinical research and healthcare domains. These past and ongoing projects in clinical trial modeling do not provide the collaborative tools needed to undertake activities in translational clinical trials, such as planning high throughput assays within a trial.

For the past two years, our research group has worked closely with trialists, scientists and developers at the Immune Tolerance Network (ITN)[4] [Rotrosen et al. 2002] to understand the day-to-day needs of authoring and managing investigator-initiated clinical trials of novel tolerance promoting therapies. The ITN also provides comprehensive mechanistic studies that complement each trial. The lifecycle management of such complex clinical trials typically involves disparate software applications facilitating activities such as trial design specification, clinical sites management, laboratory management, participants tracking, and data analysis. The lack of common nomenclature among the different sources of the tracking information and the unreliable nature of the data generation can lead to significant challenges in the operation of the clinical trials and the analysis of the research data. The applications support different but related aspects of a clinical trial, and require clinical trial data flow and knowledge exchange between the applications. The situation becomes especially critical with the need to manage complex clinical trials at various sites, and to facilitate meta-analyses on across the different trials. To support ITN's efforts, we have created and validated (1) a set of ontologies that can represent the knowledge used to execute a translational clinical trial [Shankar et al. 2006] and (2) an ontological framework that supports semantic interoperability among software applications that capture knowledge and data in ITN trials. We are now employing these methods within the recently established Human Immune Monitoring Center (HIMC)[5] at Stanford

[1] HL7: http://www.hl7.org/
[2] CDISC: http://www.cdisc.org/standards/
[3] BRIDG: http://www.bridgproject.org/
[4] ITN: http://www.immunetolerance.org/
[5] HIMC: http://imc.stanford.edu

University, which aims to provide cutting-edge basic science methods to physician investigators engaged in clinical trials and pursuing mechanistic hypotheses.
.

Epoch Clinical Trial Ontologies

A clinical trial protocol (the plan for a trial) lays out specification, implementation and data analysis details. For example, it includes the reason for undertaking the study, the number of participants that will be in the study and the recruitment process, the sites where the study will be conducted, the study drug that the participants will take, the medical tests that the participants will undergo, the data that will be collected, and the statistical analyses that will be performed on the data. We have developed Epoch, a suite of clinical trial ontologies that formally represents protocol entities relevant to the clinical trials management applications that we are supporting.

- The *clinical trial ontology* is the overarching ontology that includes references to protocol specification and operational plan. ITN has an enormous requirement on the collection and processing of specimens to support its immunological studies. The operational plan contains specifications of specimen workflow including the type of specimen containers used and the assays performed.
- The *protocol ontology* is a knowledge model of the clinical trial protocol. The main concepts represented in the protocol ontology are the protocol schema and the schedule of activities.
- The *organization ontology* provides a structure to specify study sites, clinical and core laboratories, and bio-repositories that participate in the implementation of a specific protocol.
- The *assay ontology* models characteristics of mechanistic studies relevant to immune disorders. An assay specification includes the clinical specimen that can be analyzed using that assay, and the workflow of the specimen processing at the core laboratories.
- The *labware ontology* models a laboratory catalog that mainly lists specimen containers used in the clinical trials.
- The *virtual trial data ontology* encapsulates the study data that is being collected, such as participant clinical record, specimen workflow logs and site related data.
- The *constraint expression ontology* contains formalisms for representing logical and temporal constraints found in a protocol.
- The *measurement ontology* has concepts of physical measurements such as volume and duration, and units of measurement such as milliliter and month.

The Epoch ontologies thus provide a common nomenclature and semantics required to support an integrated and consistent clinical trials management.

We have developed these ontologies in OWL (the Web Ontology Language proposed by W3C) [http://www.w3.org/TR/owl-features/] by building hierarchies of classes describing concepts in the ontologies and relating the classes to each other using properties. We use SWRL (the Semantic Web Rule Language) [http://www.w3.org/Submission/SWRL/] to specify rules that validate the constraints specified using the constraint expression ontology. Protégé [Knublauch et al. 2004] is a software tool that supports the specification and maintenance of terminologies, ontologies and knowledge-bases. We used Protégé to create the ontologies in OWL.

An Ontological Framework to Support Semantic Integration

We have built an ontology-based architecture that supports three broad types of methods that clinical trials management applications can use to interoperate.

Knowledge acquisition methods allow users to encode specific protocols and related operational elements to create the protocol knowledge base. Protocol-encoding is a labor-intensive engineering process that requires not only detailed understanding of clinical content, but also modeling formalisms. Thus, the process can be overwhelming, especially for domain experts who are not familiar with Epoch ontologies. We are building TrialWiz, a protocol authoring tool, to manage the complexity of the protocol-encoding process, and to improve efficiency in knowledge acquisition. The features provided by TrialWiz include (1) intelligent guidance in the protocol-encoding process, (2) graphical user interfaces intuitive to clinical trialists, (3) a repository of reusable knowledge on protocol components, (4) explanation and validation facilities, (5) facilities to export to different document formats and formal models, and (6) web-based collaborative authoring environment.

Knowledge sharing methods facilitate sharing the clinical trial semantics in the Epoch knowledge bases with data collection and data analysis applications. The methods employ semantic web technologies to support integration of heterogeneous applications that share the semantics of the clinical trials and not necessarily the representation formalisms. The ITN applications that we are integrating have been built by different software vendors using their proprietary information representations and with different application interface requirements. We have developed few techniques that the applications can use to obtain clinical trial knowledge relevant to their workings: 1) a SWRL-based tool that generates XML renditions of the Epoch knowledge base based on custom XML Schema, 2) a SWRL-based tool that exports portions of the knowledge base to application databases, 3) an application program

interface (api) based on Protégé-OWL api for direct access of the Epoch knowledge base and web services for remote access, and 4) a framework that maps Epoch ontologies to other clinical trial models such that the semantics in the Epoch knowledge base can be shared with applications that have been built around 'non-Epoch' models. Examples of how we use these methods with ITN applications can be found elsewhere [Shankar et al. 2006].

Ontology-database mapping methods integrate the protocol and biomedical knowledge with ITN's data repository. The repository is a relational database system that stores data related to the implementation and execution of clinical trials. The types of data include participant enrollment data, specimen shipping and receiving logs, participant visits and activities records, and clinical assessment and assay results. The mapping methods [O'Connor et al. 2007] use a schema ontology in OWL that provides a knowledge-level description of a relational schema. The ontology describes schemas in a database and associated tables together with the columns and column data types contained in each table. It also describes primary and foreign key relationships for tables in a schema. A mapping ontology then uses this schema ontology to describe how relational tables are to be mapped to concepts in Epoch's virtual trial data ontology. We have also developed the mapping software that uses the schema and mapping ontologies to transform the data in the relational databases to Epoch OWL entities. We extended our existing query engine to interact with the mapping software to retrieve the mapped OWL entities. The query engine takes SWRL queries written in terms of OWL classes, properties, and individuals and generates data requests to the mapping software. The mapping software uses the schema and mapping ontologies, and generates SQL queries to retrieve appropriate data from the database.

Discussion

The increasing complexity of clinical trials has generated an enormous requirement for knowledge and information specification at all stages of the trials, including planning, specification, implementation, and analysis. The knowledge representation and reasoning requirements borne out of the need for semantic interoperability in our clinical trial management system align well with the touted strengths of semantic technologies – uniform domain-specific semantics, flexible information models, and inference technology.

We highlight two issues in our work. First, we used OWL to specify the ontologies, and SWRL rules written in terms of concepts in these ontologies to express any constraints. With the knowledge sharing methods where we generate XML-renditions of the knowledge base or where we map the Epoch ontologies to other clinical trial models, it is not easy to export the semantics of the constraints. We are currently working on a declarative rules framework [Shankar et al. 2008] wherein constraints are specified using high level constructs in the constraints expression ontology. The constructs and their attributes can then be "assembled" as SWRL rules at a later implementation stage. Then the knowledge sharing methods can use the constructs to effectively share the semantics of the rules. The second issue is with the ontology-database mapping. We use a virtual data model to interface with the clinical trial data repository using a SWRL-based mapping methodology. At runtime, when querying for data required for executing a constraint rule, the mapping tool retrieves relevant data from the repository for the rule engine to use. We are exploring techniques to optimize the amount of data that is being retrieved, but scalability will continue to be our concern.

There are several efforts [Zhang et al 2005, Vdovjak et al 2001] from the semantic web community that propose similar ontology-based architectures to integrate distributed information resources. The ITN applications that we are integrating have been built by different software vendors using their proprietary information representations. Using semantic approaches, we are able to integrate existing software applications and databases at semantic levels so as to improve clarity, consistency and correctness in specifying clinical trials, and in acquiring and analyzing translational clinical trials data.

References

Hoffman, M.A. 2007. The genome-enabled electronic medical record. *Journal of Biomedical Informatics* 40(1): 44–46.

Knublauch, H., Fergerson, R.W., Noy, N.F. and Musen, M.A. 2004. The Protégé OWL Plugin: An open development environment for semantic web applications, Proceedings of the Third International Semantic Web Conference, 229–243.

O'Connor, M.J., Shankar, R.D., Tu, S.W., Parrish, D.B., Das, A.K., Musen, M.A. 2007. Using Semantic Web Technologies for Knowledge-Driven Querying of Biomedical Data. Proceedings of the Eleventh Conference on Artificial Intelligence in Medicine, 267-276.

Ransohoff, D.F. 2005. Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews Cancer* 5: 142–149.

Rotrosen, D., Matthews, J.B., Bluestone, J.A. 2002. The immune tolerance network: a new paradigm for developing

tolerance-inducing therapies. *Journal of Allergy and Clinical Immunology* 110(1):17–23.

Shankar, R.D., Martins, S.B., O'Connor, M.J., Parrish, D.B., Das, A.K. 2006. Epoch: an ontological framework to support clinical trials management. Proceedings of the International Workshop on Health Information and Knowledge Management, 25–32.

Shankar, R.D., Martins, S.B., O'Connor, M. J., Parrish, D.B., Das, A. K. 2008. An Ontological approach to representing and reasoning with temporal constraints in clinical trial protocols. Proceedings of the International Conference on Health Informatics. *Forthcoming*.

Vdovjak, R., Houben, G. 2001. RDF based architecture for semantic integration of heterogeneous information sources. International Workshop on Information Integration on the Web.

Weng, C., Gennari, J.H., Fridsma, D.B. 2007. User-centered semantic harmonization; A case study. *Journal of Biomedical Informatics* 40: 353–364.

Zhang, L., Gu, J. 2005. Ontology based semantic mapping architecture. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics.