



POSTER SESSION ABSTRACTS

updated 02/11/2010

Poster 01: Biological Pathways Exchange - BioPAX Level3

Nadia Anwar
Memorial Sloan-Kettering Cancer Center
New York

Abstract:

The BioPAX ontology (www.biopax.org) is a standard language for formally representing biological pathways and is available to the biological community to enable exchange and integration of pathway data. BioPAX has been a community effort spanning 7 years, culminating in the recent release of BioPAX level 3. Level 3 supports the representation of metabolic pathways, signal transduction pathways, protein-protein interaction networks, gene regulatory networks and genetic interactions. We will outline data representation in BioPAX and the use of the BioPAX ontology in integration, analysis and visualisation of pathway data, which enables efficient use and reuse of these data. We wish to highlight the successes of this community project, the core entities within the BioPAX ontology that have changed from Level 2, demonstrate example SPARQL queries across heterogeneous pathway related data, planned developments and enhancements to the ontology and finally, outline a successful use case of data integration using OWL within the PathwayCommons knowledge base.

Data exchange and integration continues to be a challenge given the complex nature of both pathway data and data sources. Biological pathways are constructs that biologists use to represent relationships between and within chains of cellular events. Metabolic pathways typically represent flow of chemical reactions, while signal transduction pathways represent the chain of molecules that are used to transmit an external signal received by a cell to deliver the response within the cell. The data is as heterogeneous as its numerous sources (pathguide.org). BioPAX was developed to address these issues and to ease the access, use, exchange and aggregation of pathway data. The BioPAX pathway ontology is defined using the Web Ontology Language, OWL, with the view of facilitating automatic processing and the integration of information held in biological pathways. The use of OWL offers significant advantages over standard data exchange strategies which usually employ XML-Schema. Since OWL can be represented in XML, standard data exchange is automatically supported and using OWL, semantic integration of pathways offers considerable benefits. Specifically, BioPAX can be used to address the problems associated with semantic heterogeneity across data sources. In data

integration, single domain models and ontologies were first applied to overcome semantic heterogeneity. In this integration architecture, the content of data sources that are to be integrated are mapped to a global ontology and queries across heterogeneous data sources are expressed in terms of the ontology. PathwayCommons uses such an architecture. Data sources providing BioPAX files are aggregated into a single resource that can be queried using a web interface or web API (pathwaycommons.org).

In addition, there are several software components developed to be used with BioPAX files. A BioPAX validator is available at biopax.org/validator. The PAXTools Java API supports programmatic access to BioPAX OWL files with export, import and analysis, including an experimental algorithm for integrating pathways based on similar interactions. At the recent BioPAX community workshop there were many groups working on pathway visualisation using BioPAX files. There are currently 9 databases that actively support and export BioPAX files and several more data providers are working towards exporting their data in BioPAX Level 3.

Poster 02: Knowledge-driven Drug Development: The Quest for a Semantic Repository of Clinical Studies

Kaushal Desai
AstraZeneca Pharmaceuticals
Wilmington, DE

Abstract:

The rising costs of drug development make it imperative for pharmaceutical companies to quickly learn from knowledge generated in clinical trials. Timely access to successful study designs and trial outcomes may deliver speed and quality improvements in decision-making at various stages of the drug development process. Utilization of information standards for extraction, integration and exploitation of structured and unstructured clinical study information could prove to be critical in this context.

This poster will describe our implementation of a semantic repository of clinical studies in a global clinical development organization. We will discuss the design of a semantic annotation platform for clinical study reports stored in large clinical document repositories. The extracted semantic annotations over clinical studies are integrated with structured information from a global trial execution database and stored in an organizational semantic repository. Beside clinical trial models the repository also incorporated and aligned existing thesauri, dictionaries and taxonomies in a syntactically coherent and semantically sound ontology. We will discuss the practical utility of our approach emphasizing the diverse search and navigation methods based on hybrid indexes over textual content, semantic annotations, ontologies and document structure as well as reasoning potential with clinical study information in the presence of a fit-for-purpose semantic environment. We will also discuss the key implementation challenges, including the

need to agree on an organization-wide model for clinical studies, integration of existing document indices with semantic annotations and structured metadata, novel co-occurrence based faceted search and navigation over extracted trial metadata; and automatic semantic annotation based on information extraction over unstructured textual content.

Poster 03: Ontology-based Approach for Representing a Personal Exposure History

Stacy Doore
University of Maine
Orono, ME

Abstract:

Objectives:

1. Develop a conceptual framework for a personal exposure history by defining key components of a domain ontology.
2. Compare a set of competency questions in terms of the types of spatial, temporal and spatio-temporal queries that can be posed to the ontology.

Motivation:

Analysis of possible relationships of long latency disease to environmental risk factors becomes complicated by the reality of changing spatial and temporal factors in the population and the environment. In order to document the location and duration of possible exposures to harmful agents over an individual's life, an analysis strategy must include a host of factors operating on disparate scales and measures. This poster lays out the framework for such a concept.

Method:

This approach uses a number of existing upper level ontologies to provide the foundational concepts for time and space. The domain level of the ontology defines a common vocabulary for researchers to share information about a person's movement over time and the connections to environmental toxic agents in multiple daily environments. The ontology and test data set were transformed into Resource Description Framework (RDF) statements and imported into an RDF store to test queries on relationships. RDF is an inherently relationship centric format which can be queried directly on its relationships using the SPARQL query language. AllegroGraph © (Version 3.2), a graph based data store was used to create and query the RDF store. Its visualization component, Gruff © (Version 1.4.1) was used to display query results. Competency questions were formulated as SPARQL queries and used to test the knowledgebase on spatial, temporal, spatial-temporal and thematic relationships.

Description:

Framework (RDF) statements and imported into an RDF store to test queries on relationships. RDF is an inherently relationship centric format which can be queried

directly on its relationships using the SPARQL query language. AllegroGraph © (Version 3.2), a graph based data store was used to create and query the RDF store. Its visualization component, Gruff © (Version 1.4.1) was used to display query results. Competency questions were formulated as SPARQL queries and used to test the knowledgebase on spatial, temporal, spatial-temporal and thematic relationships.

Results:

While many current approaches address the important elements of the ‘where’ and ‘when’ of events in a person’s life, this ontology contributes richer semantics on relationships of individuals to locations that captures the dynamics of individuals and specificity in their relationships to locations. This ontology to RDF provides a new way to evaluate environmental health risks beyond the traditional person to location layer approach used in geographic information systems. This conceptual framework contributes a unique solution to the problem of representing complex semantics associated with location environmental attributes and a person’s location history in multiple settings.

Conclusions:

- The capacity to represent exposure risk over time for individuals presents the opportunity to aggregate common locations among groups of people based on shared relationships with locations in their past (i.e. shared residence, shared workplace, school building cohort). It is possible to identify risk groups based on their relationships to specific locations.
- The conversion of the ontology into resource description framework (RDF) graphs and use of SPARQL queries demonstrates the framework’s ability to represent and query semantically explicit event-event relationships and provide machine-interpretable definitions of basic concepts and relations among them.
- SPARQL is limited in its ability to efficiently and accurately work with complex spatio-temporal queries. Further development of this conceptual framework will benefit from ongoing refinements of SPARQL’s capacity to retrieve spatio-temporal data.

Poster 04: Facilitating the Creation of Semantic Health Information Models from XML Contents

Ariel Farkash
IBM
Haifa, IL

Abstract:

Biomedical semantic interoperability is enabled by using standard exchange formats. Further constraining these formats unifies the exchanged data into a semantically unambiguous format that improves interoperability and makes operations on the data straightforward from a technological standpoint. A healthcare IT domain expert familiar with healthcare data representation methods

and standards is typically capable of creating health interoperability models and generating instances conforming to those models. A clinical domain expert, however, is mostly familiar with the data instances and terminologies and is less comfortable with representation models. Those differences in orientation and skills form a gap where the clinical domain expert cannot review and edit the models, and the healthcare IT domain expert cannot get feedback for the created models. This paper describes a solution to this fundamental problem by utilizing templates, which are standardized sets of constrained over generic standards. The solution involves generating a template model from an instance-like template skeleton (note that the reverse direction is available via conventional instance generation tools).

The HL7 v3 Reference Information Model (RIM) is used to derive consistent health information standards such as laboratory results, medications, patient care, public health, and clinical research. It is an ANSI and ISO-approved standard that provides a unified health data “language”™ to represent complex associations between entities who play roles that participate in acts. Clinical Document Architecture (CDA) is a constrained subset of the RIM that specifies terminology-encoded structure and semantics for clinical documents. These documents can be serialized to XML that conforms to a published W3C XML Schema. Yet, a CDA model is still generic in the sense that it can capture versatile clinical content ranging from discharge summaries to referral letters and operative notes. Thus, the general CDA structure is further constrained by a set of templates that are standardized by creating a Template Model.

Our approach starts with a clinical domain expert, familiar with the clinical data in its most basic XML representation. Using a common XML editor the expert can easily place the data elements in their appropriate context in order to explicitly represent the semantics of the data. Next, an annotated minimal complete instance (tagged template skeleton) is created in much the same manner. This template skeleton is, in fact, an instance that must contain all relevant metadata with cardinality of one, but instead of actual values, it will contain data annotations.

Once an initial template skeleton is ready (along with a small set of additional metadata) the clinical domain expert can use our engine to generate the full blown template model. The engine is based on UML2 library coupled with API supplied by open source tooling developed by the Eclipse OHT project. Having a template model the common instance generation mechanism may be used to generate standard instances from the data. The resulting instances can then be reviewed by the clinical domain expert allowing him to perform additional refinements to the template skeleton. This approach creates a valuable feedback cycle bridging the clinical and healthcare IT domains.

Poster 05: Using Standardized XML Models to Enable Semantic Warehousing

Ariel Farkash
IBM
Haifa, IL

Abstract:

The use of Extensible Markup Language (XML) in healthcare and life sciences (HCLS) is spreading rapidly recently. The expressive power of XML is crucial to describe the complex phenomena in HCLS. For purposes of biomedical data semantics and information exchange there is a need to standardize the XML content to support semantic interoperability.

Traditionally, information standards are being used for information exchange (e.g., messages and services) but HCLS standardized XML models can also be used to create an underlying data model for semantic warehousing. We propose an approach whereby inbound data is persisted into the XML based warehouse in its native XML format, expanding the common approach of using XML solely as a means for exchange. This makes it easier to preserve the full richness of the source information being integrated in a warehouse while surfacing up the similarities found in data sets received from multiple data sources.

Standards developing organizations such as HL7 and CEN are using XML as part of the implementation specification of their new generation of HCLS standards. These standards are developed in a model-driven approach and get translated to XML schemas. For example, the HL7 v3 Clinical Document Architecture (CDA) standard represents clinical documents which are common in healthcare, e.g., referral letters and operative notes.

The aggregation of all data pertaining to a patient could result in a longitudinal and cross-institutional patient-centric electronic health record (EHR). The CEN EHR 13606 standard is represented in UML and can be easily implemented in XML. In life sciences, there are many XML markups to represent data such as gene expression and DNA sequencing.

Representing content of both healthcare and life sciences in XML enables the fusion of mixed content for the purpose of clinical trials as well as personalized medicine where biomarkers (e.g., genetic variants) developed in clinical trials are then used to support the clinical decision process at the point of care. Such harmonization of content representation in HCLS contributes to translational medicine where discoveries in life sciences translate to better care for patients.

Using the hierarchical nature of the XML data in the warehouse makes it possible to show the commonalities among the sources on higher level nodes while placing the varied data items on lower level nodes that appropriately extend the common structures. This makes it possible to perform semantic computations which are important in many domains of HCLS (e.g., computation of clinical context such as

resolving the subject of an observation). Thus, preserving the richness of the source data is important for semantic warehousing. However, many biomedical data consumers also need customized views of data, often based on a relational schema (e.g., for optimization of analysis). To this effect data access services are used to promote certain set of items and create data marts (e.g., relational) accommodating for user-specific models.

Work based on this approach was used to support a number of use cases varying from decision support systems for HIV care to clinical research targeted at building a disease model for Essential Hypertension.

Poster 06: Teranode Fuel - A New Level of Abstraction Required

Chris McClure
Teranode Corporation
Sudbury, MA

Abstract:

As information sources become more dynamic, the lack of or delayed access to integrated data sources present new challenges for the R&D team. To address this, a large pharmaceutical applied Teranode Fuel to support research projects within their Biotherapeutics group. Utilizing standards-based semantic technology, Fuel has improved decision tracking and execution for senior managers, decreased manual data integration time for project leaders, and provided better collective intelligence across the organization.

The use case and solution described in this poster presentation showcases how semantic technologies and standards provide a potentially transformative approach to data integration in the life science industry and beyond. Specifically, the use case will detail how Fuel extends existing SharePoint and Oracle technology to create a semantic index that integrates structured and unstructured data sources, without changes to format or location.

For IT and R&D professionals, they (1) will gain a deeper understanding of the advancements in semantic technologies, and (2) how Fuel reduces data integration time and costs, while integrating structured and unstructured data, annotations and formal decisions into the searchable data system.

Poster 07: The Cross-Cutting Semantics of Maryland Virtual Patient

Sergei Nirenburg
UMBC
Baltimore, MD

Abstract

Objectives and Motivation:

Maryland Virtual Patient (MVP) is a simulation and tutoring environment, implemented as a network of human and software agents, that is being developed to support training in clinical medicine. The human user plays the role attending physician who has the opportunity to diagnose and treat virtual patients over time in open-ended, interactive simulations. Each VP is a “double agent” composed of a realistically functioning physiological side and a reasoning- and language-enabled cognitive side. The former permits the VP to undergo the physiological states and changes associated with diseases, their treatments, and even unexpected external stimuli, such as clinically counterindicated interventions by the user. The latter permits the VP to consciously experience and reason about its disease state, make decisions about its lifestyle and medical care, discuss all of these in natural language with its attending physician (the user), and learn in various ways. A virtual tutor is available to assist the user.

Method:

All intelligent functioning in MVP – from physiological simulation, to language processing, to decision-making, to learning by intelligent agents – is carried using formal meaning representations written in the metalanguage of the OntoSem ontology. The ontology, which is language-independent and unambiguous, includes not only simple descriptions of types of objects, events and the properties that link them but also detailed scripts of complex events (e.g. disease progression), knowledge of best clinical practices, clinically relevant population-level medical knowledge, and so on. Each intelligent agent has its own version of the ontology, reflecting different inventories of world knowledge, opinions, etc. Connected to each agent’s ontology are its own ontological semantic lexicon, which permits semantically-oriented language processing, and its own fact repository, or memory of object and event instances. Over the course of MVP simulations, the VP and the tutor learn: the VP learns new medical terminology (lexicon), facts about diseases, etc. (ontology), and facts about his own disease, his physician etc. (fact repository); likewise, the tutor learns about this specific patient and this specific user/physician (fact repository).

Results:

Intelligent agents in MVP are multi-purpose. While many modeling strategies might be used for any single capability, our knowledge-based strategy supports many capabilities simultaneously, thus offering an important economy of effort. In addition, the knowledge-based approach permits us to trace the functioning of agents and readily amend the models as more information becomes available or

more detail becomes necessary, thus making the entire environment indefinitely expandable.

Conclusions:

MVP permits trainees to practice on more, and more highly differentiated, cases than would typically be encountered over a short time in real clinical experience, and it offers a more challenging and realistic experience than decision-tree type training scenarios. Ontological semantic modeling has proven effective for the wide spectrum of capabilities attributed to MVP intelligent agents.

Poster 08: Advancing Child Health Research Through Harmonized Pediatric Terminology

Riki Ohira
Booz Allen Hamilton
Rockville, Maryland

Abstract:

The pediatric clinical research domain contains unique concepts that are not prevalent in clinical research focused on adults. The Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) has an ongoing effort to establish a core library of harmonized pediatric terms through stakeholder consensus. The pediatric terminology will be reviewed and vetted by working groups containing subject matter experts in pediatrics. Consistent terminology will provide clinical researchers with tools necessary to compare and aggregate data across NICHD and other NIH Institutes/Centers; clinical research portfolios, as well as across the broader research community. In light of NICHD support of a broad clinical research portfolio, which includes many populations and age groups, the Institute decided to initiate the process of harmonization with its youngest constituents: neonates (first four weeks of life) and infants (first 24 months of life). The next area for harmonization involved labor and delivery terms and concepts. This project is taking a novel approach to harmonize terminology by using normal child development as a framework. By harmonizing pediatric concepts and terms the NICHD can then take the results from studies focused on neonates and infants and combine them with other study results to learn more and extend the impact of the Institute's research investment. The NICHD is leveraging the semantic infrastructure of the National Cancer Institute (NCI) cancer Biomedical Informatics Grid (caBIG®) and open source research tools to generate harmonized pediatric terminology and associated clinical research tools for use by the pediatric community.

Poster 09: Realizing Personalized Medicine with Semantic Technology: Applied Semantic Knowledgebases (ASK ®) at work

Robert Stanley
IO Informatics, Inc.
Berkeley, CA

Abstract:

Using a customer example for application of this technology to personalized medicine - for presymptomatic detection, scoring and stratification of patients at risk of organ failure according to combined genotypic and phenotypic information – the capabilities of an Applied Semantic Knowledgebase (“ASK”) are demonstrated. Insights gained from semantically joining coherent findings despite their different methodologies allow researchers to better understand mechanistic aspects of biomarkers for organ failure at a functional level; and to apply complex screening algorithms using SPARQL and connected statistical methods for sensitive and specific patient stratification.

Using ASK makes it possible to actively screen previously disconnected, distributed datasets, to identify and stratify results - delivering applications to be used for decision making in the life science industry and in personalized medicine. Building on core data access and integration capabilities, Sentient software applies semantic patterns to create predictive network models using virtually any combination of internal experimental data and / or external published information. These patterns apply extended semantic “Visual SPARQL” query technology to build complex searches across multiple information sets. SPARQL is capable of detecting patterns within and between different data types and relationships, even if the initial datasets are not formally joined under any common database schema or data federation method. Such patterns are then placed in an Applied Semantic Knowledgebase (ASK) which is unique to a specific research focus, providing a collection of applicable to screening and decision making. Applications include hypothesis visualization, testing and refinement; target profile creation and validation; compound efficacy and promiscuity screening; toxicity profiling and detection; disease signatures; predictive clinical trials pre-screening; and patient stratification.

Poster 10: Helping Haiti - A Semantic Web Approach to Integrating Crisis Information

Simon Twigger
Medical College of Wisconsin, Milwaukee, WI.

Abstract:

Following the Haiti earthquake on February 10th, 2010, a worldwide effort to provide aid and relief sprang into action. As with any crisis information emerged about the conditions, specific needs, people in trouble, offers of help and similar.

One significant addition to this data stream in modern crises come from social media such as Twitter. Using these tools individuals on the ground can communicate directly with the rest of the world in real time. This provides a publicly visible messaging system which can be immensely valuable in providing aid and saving lives in such a crisis situation.

Twitter is a digital data stream, has a defined API and search tools and as such can be captured and analyzed using many technologies familiar to bioinformaticians. However, trying to extract actionable data from free text using software alone is a huge challenge. To address this, the EPIC group at UC Boulder had recently developed a simple hash tag syntax for Tweets called Tweak The Tweet. This was actively promoted soon after the earthquake with the net result was that more and more structured tweets began appearing. This provided the opportunity to extract useful data from the tweet-stream which could be collected and potentially acted upon. This presents a larger opportunity in that Twitter is one of many sources of information coming out of Haiti. Others include SMS text messages and reports submitted over the web to sites such as <http://haiti.usahidi.com>. Individually these reports are valuable but if they could be integrated and augmented with other data their utility may be much greater. In the biomedical arena we have been using ontologies, RDF and related semantic web approaches to address similar integration challenges so we decided to investigate its application in this situation.

To begin to define a standard set of terms an OWL ontology was created corresponding to the hash tags and significant keywords present in the the tweets. A prototype web application (<http://tweetneed.org>) was built to capture the incoming tagged tweets and provide a visualization platform and enable the conversion to RDF. We have developed parsing algorithms to extract as much data as possible from these tweets and had to address problems such as identifying duplicate information, a common situation due to 'retweets' of important messages. A prototype triple store has been created using the Talis platform and we are investigating other data sources such as reports from haiti.usahidi.com that can be RDFized and connected together. Going forward one can imagine a number of ways in which this crisis-related data could be augmented using a semantic web approach - tweets mentioning a medical condition could be matched to appropriate drugs or the sender could be directed to functioning care facilities able to treat that condition. The poster will describe the work to date and highlight some other opportunities that this type of approach might provide to assist in crisis situations.

Poster 11: OpenDMAP Information Extraction Integrating Biological Semantics and Linguistic Syntax

Helen L Johnson, Kevin Livingston, Karin Verspoor, Larry Hunter
University of Colorado Denver
Denver, CO, USA

Curated data recorded in biological databases is a critical resource for biological data analysis. This data, however, is vastly incomplete. The biomedical literature contains much information that is not represented in databases. Mining both background and novel information from literature sources is useful to biomedical research, whether the information comes in the form of extractive summaries, triples loaded into databases, or as input to more extensive systems that visualize data from an array of sources.

The OpenDMAP (Open source Direct Memory Access Parser) concept recognition system uses patterns to extract events from biomedical text. Patterns applied using OpenDMAP for extraction of various biologic interaction types, such as protein-protein interaction, phosphorylation, localization, etc., so far have largely relied on matching a continuous sequence of pattern elements including text literals, semantically typed categories, and shallow syntactic categories. Historically, the precision of OpenDMAP output has been high, but recall has been low.

Diverse resources, both syntactic and semantic, exist to improve the performance of rule-based systems like OpenDMAP. Recall lags due to the extensive variation of expression in language of simple and complex concepts alike. To address the variation of syntactic expression, dependency parse information can be added to OpenDMAP patterns, ostensibly increasing the true positive rate by matching syntactically relevant the arguments of biological predicates even if they are not sequential in the text, and additionally reducing the false positive rate by weeding out those arguments that are not syntactically viable. An increase in recall is often accompanied by a drop in precision. To address issues of precision, additional information collected from biological databases can be linked to concepts in text, allowing patterns to specify and match more precise semantic categories.

Preliminary experiments in semantic and syntactic specification shown here were performed by creating UIMA (Unified Information Management Architecture) pipelines that included many components such as tokenizers, syntactic parsers, named entity recognizers and normalizers, and OpenDMAP. Results show that addressing syntactic complexity is necessary to achieve higher recall, and that higher precision results when layering additional semantic information culled from sources external to the text.