Role of Semantic Web in Health Informatics

Satya S. Sahoo
Wright State University - Main Campus

Guo-Qiang Zhang

Amit P. Sheth
Wright State University - Main Campus, amit.sheth@wright.edu

Follow this and additional works at: http://corescholar.libraries.wright.edu/knoesis

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation

This Tutorial is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu.
Role of Semantic Web in Health Informatics


Satya S. Sahoo, GQ Zhang  Amit Sheth
Division of Medical Informatics  Kno.e.sis Center
Case Western Reserve University  Wright State University
Outline

• Semantic Web
  o Introductory Overview
• Clinical Research
  o Physio-MIMI
• Bench Research and Provenance
  o Semantic Problem Solving Environment for *T. cruzi*
• Clinical Practice
  o Active Semantic Electronic Medical Record
Semantic Web
Landscape of Health Informatics

- Clinical Research
  - Patient Care
  - Personalized Medicine
  - Drug Development
  - Privacy
  - Cost

- Bench Research

- Clinical Practice

* Images from case.edu
Challenges

• **Information Integration: Reconcile heterogeneity**
  - **Syntactic Heterogeneity:** DOB vs. Date of Birth
  - **Structural Heterogeneity:** Street + Apt + City vs. Address
  - **Semantic Heterogeneity:** Age vs. Age at time of surgery vs. Age at time of admission

• **Humans can (often) accurately interpret, but extremely difficult for machine**
  - **Role for Metadata/Contextual Information/Semantics**
Semantic Web

- Web of Linked Data
- Introduced by Berners Lee et al. as next step for Web of Documents
- Allow “machine understanding” of data,
- Create “common” models of domains using formal language - ontologies
Resource Description Framework

- Resource Description Framework – Recommended by W3C for metadata modeling [RDF]
- A standard common modeling framework – usable by humans and machine understandable
**RDF: Triple Structure, IRI, Namespace**

- **RDF Triple**
  - **Subject**: The resource that the triple is about
  - **Predicate**: The property of the subject that is described by the triple
  - **Object**: The value of the property

- **Web Addressable Resource**: Uniform Resource Locator (URL), Uniform Resource Identifier (URI), Internationalized Resource Identifier (IRI)

- **Qualified Namespace**: `http://www.w3.org/2001/XMLSchema#` as `xsd:
  - `xsd: string` instead of `http://www.w3.org/2001/XMLSchema#string`
RDF Representation

- Two types of property values in a triple
  - Web resource
  - Typed literal

- The graph model of RDF: node-arc-node is the primary representation model

- Secondary notations: Triple notation
  - `companyExample:IBM companyExample:has-Total-Employee “430,000”^^xsd:integer`
• RDF Schema: Vocabulary for describing groups of resources [RDFS]
RDF Schema

- **Property** domain (rdfs:domain) and range (rdfs:range)

  Domain: Company  
  Headquarters located in  
  Range: Geographical Location

- **Class Hierarchy/Taxonomy**: rdfs:subClassOf

  SubClass: Computer Technology Company, Banking Company, Insurance Company  
  rdfs:subClassOf  
  (Parent) Class: Company
Ontology: A Working Definition

• Ontologies are shared conceptualizations of a domain represented in a formal language*

• Ontologies in health informatics:
  o Common representation model - facilitate interoperability, integration across different projects, and enforce consistent use of terminology
  o Closely reflect domain-specific details (*domain semantics*) essential to answer end user
  o Support *reasoning* to discover implicit knowledge

* Paraphrased from Gruber, 1993
OWL2 Web Ontology Language

• A language for modeling ontologies [OWL]
• OWL2 is declarative
• An OWL2 ontology (schema) consists of:
  o Entities: Company, Person
  o Axioms: Company employs Person
  o Expressions: A Person Employed by a Company = CompanyEmployee

• Reasoning: Draw a conclusion given certain constraints are satisfied
  o RDF(S) Entailment
  o OWL2 Entailment
• **Class Disjointness**: Instance of class A cannot be instance of class B

• **Complex Classes**: Combining multiple classes with set theory operators:
  
  - **Union**: Parent = ObjectUnionOf (:Mother :Father)
  
  - **Logical negation**: UnemployedPerson = ObjectIntersectionOf (:EmployedPerson)
  
  - **Intersection**: Mother = ObjectIntersectionOf (:Parent :Woman)
OWL2 Constructs

- Property restrictions: defined over property
- Existential Quantification:
  - Parent = ObjectSomeValuesFrom (:hasChild :Person)
  - To capture incomplete knowledge
- Universal Quantification:
  - US President = objectAllValuesFrom (:hasBirthPlace United States)
- Cardinality Restriction
SPARQL: Querying Semantic Web Data

• A SPARQL query pattern composed of triples
• Triples correspond to RDF triple structure, but have variable at:
  o Subject: ?company ex:hasHeadquaterLocation ex:NewYork.
  o Object: ex:IBM ex:hasHeadquaterLocation ?location.
• Result of SPARQL query is list of values – values can replace variable in query pattern
SPARQL: Query Patterns

• An example query pattern

PREFIX ex:<http://www.eecs600.case.edu/>
SELECT ?company ?location WHERE
{?company ex:hasHeadquaterLocation ?location.}

• Query Result

<table>
<thead>
<tr>
<th>company</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>NewYork</td>
</tr>
<tr>
<td>Oracle</td>
<td>RedwoodCity</td>
</tr>
<tr>
<td>MicorosoftCorporation</td>
<td>Bellevue</td>
</tr>
</tbody>
</table>

Multiple Matches
SPARQL: Query Forms

• **SELECT**: Returns the values bound to the variables

• **CONSTRUCT**: Returns an RDF graph

• **DESCRIBE**: Returns a description (RDF graph) of a resource (e.g. IBM)
  ○ The contents of RDF graph is determined by SPARQL query processor

• **ASK**: Returns a Boolean
  ○ True
  ○ False
Semantic Web + Clinical Research Informatics = Physio-MIMI
Physio-MIMI Overview

- Physio-MIMI: Multi-Modality, Multi-Resource Environment for Physiological and Clinical Research
- NCRR-funded, multi-CTSA-site project (RFP 08-001) for providing informatics tools to clinical investigators and clinical research teams at and across CTSA institutions to enhance the collection, management and sharing of data
- Collaboration among Case Western, U Michigan, Marshfield Clinic and U Wisconsin Madison
- Use Sleep Medicine as an exemplar, but also generalizable
- Two year duration: Dec 2008 – Dec 2010
Features of Physio-MIMI

• Federated data integration environment
  – Linking existing data resources without a centralized data repository

• Query interface directly usable by clinical researchers
  – Minimize the role of the data-access middleman

• Secure and policy-compliant data access
  – Fine-grained access control, dual SSL, auditing

• Tools for curating PSGs
Data Access, Secondary Use
Measure not by the size of the database, but the number of secondary studies it supported
Query Interface – driven by access

- Visual Aggregator and Explorer (VISAGE)
- Federated, Web-based
- Driven by Domain Ontology (SDO)
- PhysioMap to connect autonomous data sources

*GQ Zhang et al.*

VISAGE screenshot
Components of VISAGE
Case Control Study Design

- Case-control is a common study design
- Used for epidemiological studies involving two cohorts, one representing the **cases** and the second representing the **controls**
- Adjusting matching ratio to improve statistical power
Example (CFS)

- Suppose we are interested in the question of whether sleep parameters (EEG) differ by obesity in age and race matched males
- **Case:** adult 55-75, male, BMI 35-50 (obese)
- **Control:** adult 55-75, male, BMI 20-30 (non-obese)
- **Matching** 1:2 on race (minimize race as a factor initially)
Adult 55-75, male, BMI 35-50
Adult 55-75, male, BMI 20-30
Set up 1:2 Matching
### 1:2 Matching Result

#### Table

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Details</th>
<th>Cases</th>
<th>Available Controls</th>
<th>Matched Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Race finding: African</td>
<td>5</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Race finding: Caucasian</td>
<td>34</td>
<td>162</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>Race finding: Unknown</td>
<td>1</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>40</strong></td>
<td><strong>194</strong></td>
<td><strong>80</strong></td>
</tr>
</tbody>
</table>

#### Case and Control Matched

- **Case**
- **Control**
- **Matched**
1:5 Matching?
1:5 Matching – CFS+SHHS

Modify Control to Include TWO data sources
Sleep Domain Ontology (SDO)

• Standardize terminology and semantics (define variations) [RO]
• Facilitate definition of data elements
• Valuable for data collection, data curation
• Data integration
• Data sharing and access
• Take advantage of progress in related areas (e.g. Gene Ontology)
• Improving data quality – provenance, reproducibility
Sleep Domain Ontology (SDO)
https://mimi.case.edu/concepts

<table>
<thead>
<tr>
<th><strong>UNIQUE NAME</strong></th>
<th><strong>LABELS</strong></th>
<th><strong>DATA TYPE</strong></th>
<th><strong>UNITS</strong></th>
<th><strong>RANGE</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>DrugOntology:abciximab</td>
<td>Abciximab</td>
<td>Abdominal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDO:AbdominalRespiratoryExcursionFinding</td>
<td>Abdominal respiratory excursion finding</td>
<td>A respiratory finding representing the amplitude of movement of the abdomen on inhalation and exhalation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDO:Absence</td>
<td>Absence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DrugOntology:acarbose</td>
<td>Acarbose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DrugOntology:acetohexamide</td>
<td>Acetohexamide</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DrugOntology:adenosine_diphosphate_receptor_inhibitor</td>
<td>Adenosine diphosphate (ADP) receptor</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sleep Domain Ontology (SDO)
https://mimi.case.edu/concepts

Concept SDO:ApneaHypopneaIndex

Qualified Name: SDO:ApneaHypopneaIndex
URI: http://mimi.case.edu/ontologies/2009/1/SDO.owl#ApneaHypopneaIndex
Labels: Apnea hypopnea index (AHI)
Status: ✓
Description: Calculation: total number of apneas and hypopneas divided by the number of hours of sleep.
Unit type: UnitsOntology:event_frequency
Units: UnitsOntology:events_per_hour
Range: 0 : 250
Data Type: XMLSchema:Integer
Formula: (sd0:NumberOfSleepApneaFindings + sd0:NumberOfSleepHypopneaFindings)/(sd0:TotalSleepTime)
Formula Concepts:
  - SDO:NumberOfSleepApneaFindings
  - SDO:NumberOfSleepHypopneaFindings
  - SDO:TotalSleepTime
Relations:
  - SDO:ApneaHypopneaIndex hasDataValue XMLSchema:Integer
  - SDO:ApneaHypopneaIndex hasUnit UnitsOntology:events_per_hour
  - SDO:ApneaHypopneaIndex http://purl.org/cpr/includes SDO:NumberOfSleepApneaFindings
  - SDO:ApneaHypopneaIndex http://purl.org/cpr/includes SDO:NumberOfSleepHypopneaFindings
  - SDO:ApneaHypopneaIndex http://purl.org/cpr/includes SDO:TotalSleepTime
  - SDO:ApneaHypopneaIndex statisticalMeasureOf SDO:Index

Lineage

... Polysomnography measure...
VISAGE Query Builder showing a data query on Parkinsonian Disorders and REM sleep behavior disorder with race demographics
Semantic Web + Provenance + Bench Research = $T. cruzi$ Semantic Problem Solving Environment
Semantic Problem Solving Environment for *T. cruzi*
Provenance in Scientific Experiments
Provenance in Scientific Experiments

- Gene Name
- Sequence Extraction
- Drug Resistant Plasmid
- 3' & 5' Region
- Knockout Construct Plasmid
- Plasmid Construction
- Transfection
- T. Cruzi sample
- Transfected Sample
- Drug Selection
- Selected Sample
- Cell Cloning
- Cloned Sample
- Gene Name
- T. Cruzi sample
- Cloned Sample

- ?
Provenance in Scientific Experiments

- Provenance from the French word “provenir” describes the lineage or history of a data entity
- For Verification and Validation of Data Integrity, Process Quality, and Trust
- Semantic Provenance Framework addresses three aspects [Prov]
  - Provenance Modeling
  - Provenance Query Infrastructure
  - Scalable Provenance System
Domain-specific Provenance ontology

- **Total Number of Classes**: 118
- **DL Expressivity**: $\text{ALCHQ(D)}$

### Key Concepts
- **has_agent**
- **has_participant**
- **has_temporal_parameter**
- **has_input_value**
- **has_parameter**
Provenance Query Classification

Classified Provenance Queries into Three Categories

• Type 1: Querying for Provenance Metadata
  o Example: Which gene was used create the cloned sample with ID = 66?

• Type 2: Querying for Specific Data Set
  o Example: Find all knockout construct plasmids created by researcher Michelle using “Hygromycin” drug resistant plasmid between April 25, 2008 and August 15, 2008

• Type 3: Operations on Provenance Metadata
  o Example: Were the two cloned samples 65 and 46 prepared under similar conditions – compare the associated provenance information
Provenance Query Operators

Four Query Operators – based on Query Classification

• \textit{provenance()} – Closure operation, returns the complete set of provenance metadata for input data entity

• \textit{provenance\_context()} - Given set of constraints defined on provenance, retrieves datasets that satisfy constraints

• \textit{provenance\_compare()} - adapt the RDF graph equivalence definition

• \textit{provenance\_merge()} - Two sets of provenance information are combined using the RDF graph merge
Answering Provenance Queries using \textit{provenance} () Operator
Implementation: Provenance Query Engine

- **Three modules:**
  - Query Composer
  - Transitive closure
  - Query Optimizer

- **Deployable over a RDF store with support for reasoning**
Application in *T. cruzi* SPSE Project

- Provenance tracking for gene knockout, strain creation, proteomics, microarray experiments
- Part of the Parasite Knowledge Repository [BKR]
W3C Provenance Working Group

• Define a “provenance interchange language for publishing and accessing provenance”

• Three working drafts:
  o PROV-Data Model: A conceptual model for provenance representation
  o PROV-Ontology: An OWL ontology for provenance representation
  o PROV-Access and Query: A framework to query and retrieve provenance on the Web
Semantic Web + Clinical Practice Informatics = Active Semantic Electronic Medical Record (ASEMR)
Semantic Web application in use

In daily use at Athens Heart Center

- 28 person staff
  - Interventional Cardiologists
  - Electrophysiology Cardiologists
- Deployed since January 2006
- 40-60 patients seen daily
- 3000+ active patients
- Serves a population of 250,000 people
Information Overload in Clinical Practice

• New drugs added to market
  – Adds interactions with current drugs
  – Changes possible procedures to treat an illness

• Insurance Coverage's Change
  – Insurance may pay for drug X but not drug Y even though drug X and Y are equivalent
  – Patient may need a certain diagnosis before some expensive test are run

• Physicians need a system to keep track of ever changing landscape
System though out the practice
System though out the practice
System though out the practice
System though out the practice
Active Semantic Document (ASD)

A document (typically in XML) with the following features:

• **Semantic annotations**
  – Linking entities found in a document to ontology
  – Linking terms to a specialized lexicon [TR]

• **Actionable information**
  – Rules over semantic annotations
  – Violated rules can modify the appearance of the document (Show an alert)
Active Semantic Patient Record

• An application of ASD
• Three Ontologies
  – Practice
    Information about practice such as patient/physician data
  – Drug
    Information about drugs, interaction, formularies, etc.
  – ICD/CPT
    Describes the relationships between CPT and ICD codes
• Medical Records in XML created from database
Practice Ontology Hierarchy
(showing is-a relationships)
Drug Ontology Hierarchy
(showing is-a relationships)
Drug Ontology showing neighborhood of PrescriptionDrug concept
Part of Procedure/Diagnosis/ICD9/CPT Ontology

- **specificity**
- **diagnosis** maps_to_diagnosis **procedure**
- **maps_to_procedure**
Semantic Technologies in Use

• Semantic Web: OWL, RDF/RDQL, Jena
  – OWL (constraints useful for data consistency), RDF
  – Rules are expressed as RDQL
  – REST Based Web Services: from server side

• Web 2.0: client makes AJAX calls to ontology, also auto complete

Problem:

• Jena main memory- large memory footprint, future scalability challenge

• Using Jena’s persistent model (MySQL) noticeably slower
Benefits: Athens Heart Center Practice Growth

![Graph showing appointments growth from 2003 to 2006 by month]

- **2003** (△)
- **2004** (···)
- **2005** (●)
- **2006** (—)
Chart Completion before the preliminary deployment of the ASMER

![Chart showing chart completion over months from January 2004 to July 2005. The chart compares 'Same Day' (light blue bars) and 'Back Log' (dark red bars) charts across different months.](chart.png)
Chart Completion after the preliminary deployment of the ASMER
Benefits of current system

• **Error prevention (drug interactions, allergy)**
  – Patient care
  – Insurance

• **Decision Support (formulary, billing)**
  – Patient satisfaction
  – Reimbursement

• **Efficiency/time**
  – Real-time chart completion
  – “semantic” and automated linking with billing
Demo

On-line demo of Active Semantic Electronic Medical Record

deployed and in use at Athens Heart Center
Challenges, Opportunities, and Future Direction
Conclusions

Benefits of SW in Health Informatics:

• RDF a “universal” data model; Application-purpose agnostic (clinical care vs research)

• Integration “ready,” supporting distributed query out of box

• Semantic interoperability addressed at root level

• Better support of user interfaces for data capture, data query, data integration

• Scalability demonstrated
Challenges and Future Directions

• Design and implementation of health information systems with RDF as primary data store from ground up
• User-friendly graphical query interface on top of SPARQL
• Managing Protected Health Information (PHI) e.g. data encryption “at rest” for RDF store
• From retrospective annotation of data (with ontology) to prospective annotation of data: ontology-driven data capture with annotation happening at the point of primary source (eliminating the need to annotate data retrospectively)
• Let ontology drive “everything”
References


• [Physio-MIMI]: http://physiomimi.case.edu


References 2


• T.cruzi project web site: http://wiki.knoesis.org/index.php/Trykipedia
Acknowledgements

• Collaborators:
  - Susan Redline, Remo Mueller, and other members of Physio-MIMI team
  - Rick Tarleton, Todd Manning, Priti Parikh and other members of the T.cruzi SPSE team
  - Dr. S. Agrawal and other members at the Athens Heart Center, GA

• NIH Support: UL1-RR024989, UL1-RR024989-05S, NCRR-94681DBS78, NS076965, and 1R01HL087795