Semantic Web for Health Care and Biomedical Informatics

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

Follow this and additional works at: https://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 Presentation 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 library-corescholar@wright.edu.
Semantic Web for Health Care and Biomedical Informatics

Keynote at
NSF Biomed Web Workshop, December 4-5, 2007

Amit P. Sheth
amit.sheth@wright.edu

Thanks Pablo Mendes, Satya Sahoo and Kno.e.sis team;
Collaborators at Athens Heart Center (Dr. Agrawal), NLM (Olivier Bodenreider), CCRC, UGA (Will York), CCHMC (Bruce Aronow)
Outline

• Semantic Web – very brief intro
• Scenarios to demonstrate the applications and benefit of semantic web technologies
  – Health care
  – Biomedical Research
Biomedical Informatics...  

...needs a connection

Hypothesis Validation  
Experiment design  
Predictions  
Personalized medicine

Semantic Web research aims at providing this connection!

More advanced capabilities for search, integration, analysis, linking to new insights and discoveries!

Etiology  
Pathogenesis  
Clinical findings  
Diagnosis  
Prognosis  
Treatment

Genbank  
Uniprot

Genome  
Transcriptome  
Proteome  
Metabolome  
Physiome  ...
ome

Pubmed  
Clinical Trials.gov

Biomedical Informatics  
Medical Informatics  
Bioinformatics
Evolution of the Web

Web of pages
- text, manually created links
- extensive navigation

Web of databases
- dynamically generated pages
- web query interfaces

Web of services
- data = service = data, mashups
- ubiquitous computing

Web of people
- social networks, user-created content
- GeneRIF, Connotea

Web as an oracle / assistant / partner
- “ask to the Web”
- using semantics to leverage
text + data + services + people
• **Ontology**: Agreement with Common Vocabulary & Domain Knowledge; Schema + Knowledge base

• **Semantic Annotation (metadata Extraction)**: Manual, Semi-automatic (automatic with human verification), Automatic

• **Reasoning/computation**: semantics enabled search, integration, complex queries, analysis (paths, subgraph), pattern finding, mining, hypothesis validation, discovery, visualization
Maturing capabilities and ongoing research

- Text mining: Entity recognition, Relationship extraction
- Integrating text, experimental data, curated and multimedia data
- Clinical and Scientific Workflows with semantic web services
- Hypothesis driven retrieval of scientific literature, Undiscovered public knowledge
Metadata and Ontology: Primary Semantic Web enablers

Shallow semantics

Deep semantics

Types of Metadata and Annotations

Ontology
(Example: Anatomy, Diagnostics, ...)

Semantic Metadata
(Example ontology-driven metadata:
Region: Upper Abdomen
Organ: Liver
Pathological Structure: Abscess, Abscess located in Liver)

Structural Metadata
(document structure: DTDs, XSL
clustering and similarity processing: concept extraction)

Syntactic Metadata
(language, format, document length, creation date, source,
audio bit rate, encryption, affiliation, date last reviewed, authorization, ...)

Data
(Structured, semi-structured and unstructured)
The Semantic Web: XML, RDF & Ontology

Characteristics of Semantic Web

- Easy to Understand
- Convertible
- Can be Secured
- Self Describing
- Issued by a Trusted Authority
Many ontologies exist

<table>
<thead>
<tr>
<th>Ontology Name</th>
<th>Identifier</th>
<th>URI</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESH</td>
<td>MESH</td>
<td>mesh.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Microarray experimental conditions</td>
<td>MO</td>
<td>MGEDOntology.owl</td>
<td>OWL</td>
</tr>
<tr>
<td>Molecular function</td>
<td>GO</td>
<td>gene ontology.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Molecule role (INOH Protein name/family name ontology)</td>
<td>IMR</td>
<td>molecule role.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Mosquito gross anatomy</td>
<td>TIGMA</td>
<td>mosquito anatomy.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Mouse adult gross anatomy</td>
<td>MA</td>
<td>adult mouse anatomy.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Mouse gross anatomy and development</td>
<td>EMAF</td>
<td>EMAF.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Mouse pathology</td>
<td>MPATH</td>
<td>mouse pathology.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Multiple alignment</td>
<td>RO</td>
<td>mao.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>NCI Thesaurus</td>
<td>NCIT</td>
<td>EYS ftp site</td>
<td>OWL</td>
</tr>
<tr>
<td>NMR-instrument specific component of metabolomics investigations</td>
<td>MSI</td>
<td>mnr.owl</td>
<td>OWL</td>
</tr>
<tr>
<td>OBO relationship types</td>
<td>OBO_REL</td>
<td>relationship.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Pathway ontology</td>
<td>PW</td>
<td>pathway.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>PATO</td>
<td>PATO</td>
<td>quality.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Physico-chemical methods and properties</td>
<td>FIX</td>
<td>fix.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Physico-chemical process</td>
<td>REX</td>
<td>rex.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Plant environmental conditions</td>
<td>EO</td>
<td>environment ontology.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Plant growth and developmental stage</td>
<td>PO</td>
<td>po_temporal.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Plant structure</td>
<td>PO</td>
<td>po_anatomy.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Plasmodium life cycle</td>
<td>PLO</td>
<td>PLOontology</td>
<td>GO</td>
</tr>
<tr>
<td>Protein covalent bond</td>
<td>[none]</td>
<td>[none]</td>
<td>[none]</td>
</tr>
<tr>
<td>Protein domain</td>
<td>IPR</td>
<td>InterPro FTP directory</td>
<td>XML</td>
</tr>
<tr>
<td>Protein modification</td>
<td>MOD</td>
<td>psi-mod.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Protein-protein interaction</td>
<td>MI</td>
<td>psi-mi.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Proteomics data and process provenance</td>
<td>ProPreO</td>
<td>ProPreO.owl</td>
<td>OWL</td>
</tr>
<tr>
<td>Sequence types and features</td>
<td>SO</td>
<td>so.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Systems Biology</td>
<td>SBO</td>
<td>sbo.obo</td>
<td>OBO</td>
</tr>
</tbody>
</table>
Drug Ontology Hierarchy
(showing is-a relationships)
N-Glycosylation metabolic pathway

GNT-I attaches GlcNAc at position 2

\[
\text{UDP-N-acetyl-D-glucosamine + \alpha-D-Mannosyl-1,3-(R1)-\beta-D-mannosyl-R2} \rightleftharpoons \text{UDP + N-Acetyl-\beta-D-glucosaminyl-1,2-\alpha-D-mannosyl-1,3-(R1)-\beta-D-mannosyl-R2}
\]

GNT-V attaches GlcNAc at position 6

\[
\text{UDP-N-acetyl-D-glucosamine} + \text{G00020} \rightleftharpoons \text{UDP + G00021}
\]
Opportunity: exploiting clinical and biomedical data

Collection of (heterogeneous) Documents
- text
- XML
- HTML pages
- Databases
- binary

Scientific Literature
- PubMed
- 300 Documents Published Online each day

Health Information Services
- Elsevier
- iConsult

User-contributed Content (Informal)
- GeneRifs

NCBI Public Datasets
- Genome, Protein DBs
- new sequences daily

Clinical Data
- Personal health history

Laboratory Data
- Lab tests, RTPCR, Mass spec

Search, browsing, complex query, integration, workflow, analysis, hypothesis validation, decision support.
Scenario 1:

- Status: In use **today**
- Where: Athens Heart Center
- What: Use of semantic Web technologies for *clinical decision support*
Operational since January 2006

Knowledge Enabled Information and Services Science
Goals:

- Increase efficiency with decision support
  - formulary, billing, reimbursement
  - real time chart completion
  - automated linking with billing
- Reduce Errors, Improve Patient Satisfaction & Reporting
  - drug interactions, allergy, insurance
- Improve Profitability

Technologies:

- Ontologies, semantic annotations & rules
- Service Oriented Architecture

Thanks -- Dr. Agrawal, Dr. Wingeth, and others. *ISWC2006 paper*
Demonstration
ASMER Efficiency

Chart Completion **before** the preliminary deployment

Chart Completion **after** the preliminary deployment
Scenario 2:

- **Status:** Demonstration
- **Where:** W3C Health Care and Life Sciences (HCLS) interest group
- **What:** Using semantic web to aggregate and query data about Alzheimer’s

- http://www.w3.org/2001/sw/hcls/
Scenario 2: Scientific Data Sets for Alzheimer’s

- Integration and analysis of heterogeneous data sets
  - Hypothesis, Genome, Pathways, Molecular Properties, Disease, etc.

![Diagram with various data sets and connections]
Allen Brain Atlas Gene Expression Results

3. entrez-gene-record filtered from 26 originally (reset)

sorted by: hasName; then by ...
• grouped as sorted • show duplicates

56298
Entrez-Gene 56298
ADP-ribosylation factor-like 6 interacting protein 2

Transcript Region
NC_000003.4

Genomic Context

433667
Entrez-Gene 433667
ankyrin repeat domain 13c

Transcript Region
NC_000009.4

Genomic Context

Knowledge Enabled Information and Services Science
Scenario 3

- **Status:** Completed research
- **Where:** NIH
- **What:** Understanding the genetic basis of nicotine dependence. Integrate gene and pathway information and show how three complex biological queries can be answered by the integrated knowledge base.
- **How:** Semantic Web technologies (especially RDF, OWL, and SPARQL) support information integration and make it easy to create semantic mashups (semantically integrated resources).
Motivation

• NIDA study on nicotine dependency
• List of candidate genes in humans
• Analysis objectives include:
  o Find interactions between genes
  o Identification of active genes – maximum number of pathways
  o Identification of genes based on anatomical locations
• Requires integration of genome and biological pathway information
Genome and pathway information integration

- Reactome
  - pathway
  - protein
  - PMID

- KEGG
  - pathway
  - protein
  - PMID

- HumanCyc
  - pathway
  - protein
  - PMID

- Entrez Gene

- GeneOntology
  - GO ID

- HomoloGene
  - HomoloGene ID
Deductive Reasoning

RULE: given that two genes interact with each other, given certain number of parameters being met, we can assert that the gene products also interact with each other.

IF (x have_common_pathway y) AND (x rdf:type gene) AND (y rdf:type gene) AND (x has_product m) AND (y has_product n) AND (m rdf:type gene_product) AND (n rdf:type gene_product) THEN (m \? n)
Scenario 4

- **Status:** Completed research
- **Where:** NIH
- **What:** queries across integrated data sources
  - Enriching data with ontologies for integration, querying, and automation
  - Ontologies beyond vocabularies: the power of relationships
Use data to test hypothesis

Link between glycosyltransferase activity and congenital muscular dystrophy?

Adapted from: Olivier Bodenreider, presentation at HCLS Workshop, WWW07
In a Web pages world…

Adapted from: Olivier Bodenreider, presentation at HCLS Workshop, WWW07

Congenital muscular dystrophy, type 1D

(GeneID: 9215)

has_associated_disease

Muscular dystrophy, congenital, type 1D
MIM: 608840

has_molecular_function

Acetylglucosaminyltransferase activity

Function

acetylglucosaminyltransferase activity

Process

N-acetylglucosamine metabolic process
carbohydrate biosynthetic process
glycosphingolipid biosynthetic process
muscle maintenance
protein amino acid glycosylation

GeneOntology
With the semantically enhanced data

SELECT DISTINCT ?t ?g ?d  {
    ?t is_a GO:0016757 .
    ?g has_molecular_function ?t .
    ?g has_associated_phenotype ?b2 .
    ?b2 has_textual_description ?d .
    FILTER (?d, "muscular dystrophy", "i") .
    FILTER (?d, "congenital", "i")
} 

From medinfo paper.
Adapted from: Olivier Bodenreider, presentation at HCLS Workshop, WWW07
Scenario 5

- **Status**: Research prototype and in progress
  - Workflow with Semantic Annotation of Experimental Data already in use

- **Where**: UGA

- **What**:
  - Knowledge driven query formulation
  - Semantic Problem Solving Environment (PSE) for *Trypanosoma cruzi* (Chagas Disease)
Knowledge driven query formulation

**Build your query**

Build Your Query | Advanced (SPARQL)

By navigating through the ontology schema (i.e. the definition of the possible types and interconnections available in the knowledge base), the system will guide you throughout the process of posing a question in an intuitive way, e.g. "Gene -> codes for -> Protein -> expressed in -> Epimastigote (Life Cycle Stage)".

Complex queries can also include:
- on-the-fly Web services execution to retrieve additional data
- inference rules to make implicit knowledge explicit
T. Cruzi PSE Query Interface

Build your query to the database

By navigating through the ontology schema (i.e., the definition of the knowledge base), the system will guide you throughout the process of e.g., "Gene -> codes for -> Protein -> expressed in -> Epimastigote (Life Cycle Stage)."

Figure 4: Semantic annotation of ms scientific data
**N-Glycosylation Process (NGP)**

1. **Cell Culture**
   - Extract
   - Proteolysis

2. **Glycoprotein Fraction**
   - Separation technique I

3. **Glycopeptides Fraction**
   - PNGase

4. **Glycopeptides Fraction**
   - n x m

5. **Peptide Fraction**
   - Separation technique II
   - Mass spectrometry

6. **Peptide Fraction**
   - n x m

**Data Reduction**
- Data reduction
- Peptide identification
- Binning

**Signal Integration**
- Knowledge Enabled Information and Services Science

**Peptide Fraction**
- ms data
- ms peaklist

**Glycopeptide Identification and Quantification**
- N-dimensional array

**ms/ms data**
- ms/ms peaklist
- Peptide list

**Data Correlation**
- Signal integration
Semantic Web Process to incorporate provenance

Biological Sample Analysis by MS/MS → Raw Data to Standard Format → Data Pre-process → DB Search (Mascot/Sequest) → Results Post-process (ProValt)

Semantic Annotation Applications

Raw Data | Standard Format Data | Filtered Data | Search Results | Final Output

Storage

Biological Information
### Mass Spectrometry (MS) Data

<table>
<thead>
<tr>
<th>Parent Ion ( m/z )</th>
<th>Fragment Ion ( m/z )</th>
<th>Fragment Ion Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>830.9570</td>
<td>580.2985</td>
<td>0.3592</td>
</tr>
<tr>
<td>194.9604</td>
<td>688.3214</td>
<td>0.2526</td>
</tr>
<tr>
<td>779.4759</td>
<td>779.4759</td>
<td>38.4939</td>
</tr>
<tr>
<td>784.3607</td>
<td>784.3607</td>
<td>21.7736</td>
</tr>
<tr>
<td>1543.7476</td>
<td>1543.7476</td>
<td>1.3822</td>
</tr>
<tr>
<td>1544.7595</td>
<td>1544.7595</td>
<td>2.9977</td>
</tr>
<tr>
<td>1562.8113</td>
<td>1562.8113</td>
<td>37.4790</td>
</tr>
<tr>
<td>1660.7776</td>
<td>1660.7776</td>
<td>476.5043</td>
</tr>
</tbody>
</table>

- **Parent Ion Charge:** 2
- **Parent Ion Abundance:** 38.4939
- **Fragment Ion Abundance:** 21.7736
<ms-ms_peak_list>
  <parameter
    instrument="micromass_QTOF_2_quadropole_time_of_flight_mass_spectrometer"
    mode="ms-ms"/>
  <parent_ion m-z="830.9570" abundance="194.9604" z="2"/>
    <fragment_ion m-z="580.2985" abundance="0.3592"/>
    <fragment_ion m-z="688.3214" abundance="0.2526"/>
    <fragment_ion m-z="779.4759" abundance="38.4939"/>
    <fragment_ion m-z="784.3607" abundance="21.7736"/>
    <fragment_ion m-z="1543.7476" abundance="1.3822"/>
    <fragment_ion m-z="1544.7595" abundance="2.9977"/>
    <fragment_ion m-z="1562.8113" abundance="37.4790"/>
    <fragment_ion m-z="1660.7776" abundance="476.5043"/>
</ms-ms_peak_list>
Scenario 6

- When: Research in progress
- Where: Athens Heart Center and Cincinatti Children’s Hospital Medical Center
- What: scientific literature mining
  - Dealing with unstructured information
  - Extracting knowledge from text
  - Complex entity recognition
  - Relationship extraction
Heart Failure Clinical Pathway

At Risk for Heart Failure

STAGE A
- At high risk for HF but without structural heart disease or symptoms of HF.
  - Hypertension
  - Atherosclerotic disease
  - Diabetes
  - Obesity
  - Metabolic syndrome
  - Patients using cardiotoxins
  - With FHx CM

STAGE B
- Structural heart disease but without signs or symptoms of HF.
  - Previous MI
  - LV remodeling including LVHa and low EF
  - Asymptomatic valvular disease

STAGE C
- Structural heart disease with prior symptom
  - Patient known str heart dise and shortness breath an reduced e tolerance

THERAPY

GOALS
- Treat hypertension
- Encourage smoking cessation
- Treat lipid disorders
- Encourage regular exercise
- Discourage alcohol intake, illicit drug use
- Control metabolic syndrome

DRUGS
- ACEI or ARB in appropriate patients
- Beta-blockers in appropriate patients
- Biventricular pacing
- Implantable defibrillators

Disease causes

Angiotension Receptor Blocker (ARB)

Ontology: A Framework for Schema-Driven Relationship Discovery from Unstructured Text, Ramakrishnan, et. al., ISWC 2006, LNCS 4273, pp. 583-596
Today, the Food and Drug Administration (FDA) is announcing to voluntarily withdraw Bextra
(from the market. Pfizer has agreed to of Bextra in the further discussions with the agency.

This request is based on:
- Reports of serious and potentially life-threatening skin reactions, including deaths, in patients using Bextra. The risk of these reactions in individual patients is unpredictable, occurring in patients with and without a prior history of sulfal allergy.

Date created: April 7, 2005
• Two technical challenges
  – Text mining
  – Workflow adaptation
Diabetes mellitus adversely affects the outcomes in patients with myocardial infarction (MI), due in part to the exacerbation of left ventricular (LV) remodeling. Although angiotensin II type 1 receptor blocker (ARB) has been demonstrated to be effective in the treatment of heart failure, information about the potential benefits of ARB on advanced LV failure associated with diabetes is lacking. To induce diabetes, male mice were injected intraperitoneally with streptozotocin (200 mg/kg). At 2 weeks, anterior MI was created by ligating the left coronary artery. These animals received treatment with olmesartan (0.1 mg/kg/day; n = 50) or vehicle (n = 51) for 4 weeks. Diabetes worsened the survival and exaggerated echocardiographic LV dilatation and dysfunction in MI. Treatment of diabetic MI mice with olmesartan significantly improved the survival rate (42% versus 27%, P < 0.05) without affecting blood glucose, arterial blood pressure, or infarct size. It also attenuated LV dysfunction in diabetic MI. Likewise, olmesartan attenuated myocyte hypertrophy, interstitial fibrosis, and the number of apoptotic cells in the noninfarcted LV from diabetic MI. Post-MI LV remodeling and failure in diabetes were ameliorated by ARB, providing further evidence that angiotensin II plays a pivotal role in the exacerbated heart failure after diabetic MI.

Angiotensin II type 1 receptor blocker attenuates exacerbated left ventricular remodeling and failure in diabetes-associated myocardial infarction. Matsusaka H, et. al.

Knowledge Enabled Information and Services Science
Problem – Extracting relationships between MeSH terms from PubMed
Background knowledge used

- **UMLS** – A high level schema of the biomedical domain
  - 136 classes and 49 relationships
  - Synonyms of all relationship – using variant lookup (tools from NLM)
  - 49 relationship + their synonyms = ~350 mostly verbs

- **MeSH**
  - 22,000+ topics organized as a forest of 16 trees
  - Used to query PubMed

- **PubMed**
  - Over 16 million abstract
  - Abstracts annotated with one or more MeSH terms
Method – Parse Sentences in PubMed

[1254239-1] An excessive endogenous or exogenous stimulation by estrogen induces adenomatous hyperplasia of the endometrium.

SS-Tagger (University of Tokyo)
SS-Parser (University of Tokyo)

• Entities (MeSH terms) in sentences occur in modified forms
  (TOP (S (NP (NP (DT An) (JJ excessive) (ADJP (JJ endogenous) (CC or) (JJ exogenous)) (NN stimulation)) (PP (IN by) (NP (NN estrogen))))) (VP (VBZ induces) (NP (NP (JJ adenomatous) (NN hyperplasia)) (PP (IN of) (NP (DT the) (NN endometrium))))))

• “adenomatous” modifies “hyperplasia”
  • An excessive endogenous or exogenous stimulation modifies estrogen

• Entities can also occur as composites of 2 or more other entities
  • “adenomatous hyperplasia” and “endometrium” occur as “adenomatous hyperplasia of the endometrium”
An excessive endogenous or exogenous stimulation by estrogen induces adenomatous hyperplasia of the endometrium.
• What can we do with the extracted knowledge?
• Semantic browser demo
Evaluating hypotheses

Keyword query: Migraine[MH] + Magnesium[MH]
Workflow Adaptation: Why and How

• Volatile nature of execution environments
  – May have an impact on multiple activities/ tasks in the workflow

• HF Pathway
  – New information about diseases, drugs becomes available
  – Affects treatment plans, drug-drug interactions

• Need to incorporate the new knowledge into execution
  – capture the constraints and relationships between different tasks activities
New knowledge about treatment found during the execution of the pathway

New knowledge about drugs, drug drug interactions
Workflow Adaptation: How

• Decision theoretic approaches
  – Markov Decision Processes

• Given the state S of the workflow when an event E occurs
  – What is the optimal path to a goal state G
  – Greedy approaches rely on local optimization
    • Need to choose actions based on optimality across the entire horizon, not just the current best action
  – Model the horizon and use MDP to find the best path to a goal state
Conclusion

• semantic web technologies can help with:
  – Fusion of data: semi-structured, structured, experimental, literature, multimedia
  – Analysis and mining of data, extraction, annotation, capture provenance of data through annotation, workflows with SWS
  – Querying of data at different levels of granularity, complex queries, knowledge-driven query interface
  – Perform inference across data sets
Take home points

- Shift of paradigm: from browsing to querying
- Machine understanding:
  - extracting knowledge from text
  - Inference, software interoperation
- Semantic-enabled interfaces towards hypothesis validation
References


• Demos at: http://knoesis.wright.edu/library/demos/