Use Cases, Data & Ontologies for Pharma & Translational Medicine

Elgar Pichler
HCLSIG TMO & LODD Team Members

CSHALS 2010, Cambridge, MA
February 24, 2010
Outline

- Questions & Problems
- LODD & TMO
- TMO Development
- Data & Tools
- Example
- Summary
Questions & Problems

Aspirin – nothing new, right?

New findings every day.

How does this affect the use of a drug? How does it affect me?
Questions & Problems
The Drug Development Pipeline

• The road is long, and costly.
• How do we contain costs and develop better drugs?

http://www.nature.com/horizon/chemicalspace/background/odyssey.html
LODD & TMO

• LODD
  – Focuses on linking various sources of drug data – ranging from data describing the impact of drugs on gene expression, through to clinical trial results – to answer interesting scientific and business questions.

• TMO
  – Focuses on the development of a high level patient-centric ontology for the pharmaceutical industry. The ontology should enable silos in discovery research, hypothesis management, experimental studies, compounds, formulation, drug development, market size, competitive data, population data, etc. to be brought together. This would enable scientists to answer new questions, and to answer existing scientific questions more quickly. This will help pharmaceutical companies to model patient-centric information, which is essential for the tailoring of drugs, and for early detection of compounds that may have sub-optimal safety profiles. The ontology should link to existing publicly available domain ontologies.
LODD data in the Linked Data cloud are represented in dark gray. Collectively, the data sets consist (August 2009) of over 8 million RDF triples, which are interlinked by more than 370,000 RDF links.
LODD

Aspirin?

Aspirin, also known as acetylsalicylic acid (abbreviated ASA), is a salicylate drug, which relieves pain and fever, and as an anti-inflammatory medication. By inhibiting the production of thromboxanes, which under normal circumstances binds platelet molecules together to create a patch over damage of the walls within blood vessels, because the platelet patch can become too large and also block blood flow, locally and downstream, aspirin is also used long-term, at low doses, to help prevent heart attacks, strokes, and blood clot formation in people at high risk for developing blood clots. It has also been established that low doses of aspirin may be given immediately after a heart attack to reduce the risk of another heart attack or of the death of cardiac tissue. The main undesirable side effects of aspirin are gastrointestinal ulcers, stomach bleeding, and ulcers, especially in higher doses. In children and adolescents, aspirin is no longer used to control flu-like symptoms or the symptoms of chickenpox or other viral illnesses, because of the risk of Reye’s syndrome.

Easy to query for drug related information.
LODD

• home:
  – http://esw.w3.org/topic/HCLSIG/LODD

• data sources (with SPARQL endpoints list):
  – http://esw.w3.org/topic/HCLSIG/LODD/Data
  – http://hcls.deri.org/sparql

• examples
  – http://www4.wiwiss.fu-berlin.de/lodd/topquestions/
TMO Development

Concept Identification via Use Cases

• Process:
  – work out use cases
  – identify used concepts
  – map concepts to other ontologies/vocabularies
  – align with Basic Formal Ontology (BFO)
  – refine and start over again
TMO Development
Concept Identification via Use Cases

Example

(see http://esw.w3.org/topic/HCLSIG/PharmaOntology/UseCases):


[...]
TMO Development
Mapping to Other Ontologies/Vocabularies

NCBO

UMLS
### TMO Mapping to Other Ontologies/Vocabularies

#### Mapping examples:

<table>
<thead>
<tr>
<th>TMO class</th>
<th>Classes in other ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>pharmaceutical product (TMO_0002)</td>
<td>NCIt:Finished_Pharmaceutical_Product, UMLS:C1708062</td>
</tr>
<tr>
<td>target (TMO_0006)</td>
<td>NCIt:Target, OCRe:research2:target, UMLS:C1521840</td>
</tr>
<tr>
<td>institution (TMO_0025)</td>
<td>ACGT:Institution, BIRNLex:2085, LNC:LP76237-4, NCIt:Institution, SNOMEDCT:385437003, UMLS:C1272753</td>
</tr>
<tr>
<td>intervention (TMO_0030)</td>
<td>ClinicalTrialOntology:prtont:PeriodType_5, NCIt:Intervention, OCRe:research2:Intervention</td>
</tr>
<tr>
<td>clinical trial (TMO_0032)</td>
<td>HL7V3.0:CLNTRL, MSH:D016430, NCIt:Clinical_Trial, SNOMEDCT:110465008</td>
</tr>
<tr>
<td>disease (TMO_0047)</td>
<td>ACGT:Disease, BIRNLex:11013, DOID:4, GRO:Disease, LNC:LP21006-9, MSH:D004194, NCIt:Disease_or_Disorder, NDFRT:C2140, OBI:0000155</td>
</tr>
</tbody>
</table>
TMO Development
Alignment with BFO
<table>
<thead>
<tr>
<th>Name</th>
<th>Topic</th>
<th>Short Description</th>
<th>Size</th>
<th>LODD</th>
<th>TMO</th>
</tr>
</thead>
<tbody>
<tr>
<td>DailyMed</td>
<td>Drugs</td>
<td>dailymed.nlm.nih.gov provides information about approved prescription drugs, includes FDA approved labels (package inserts).</td>
<td>164,276 triples; 4,039 drugs</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>DBpedia</td>
<td>Drugs / Diseases / Proteins</td>
<td>RDF data about 2.49 million things that has been extracted from Wikipedia.</td>
<td>218M triples; 2,300 drugs; 2,200 proteins</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Diagnostic Data</td>
<td>Disease / Diagnosis</td>
<td>AD specific diagnostic data extracted from a paper by DuBois et al (2007).</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Diseasome</td>
<td>Diseases / Genes</td>
<td>Diseasome describes characteristics of disorders and disease genes linked by known disorder–gene associations.</td>
<td>91,182 triples; 2,600 genes</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>DrugBank</td>
<td>Drugs</td>
<td>Drugbank.ca provides drug (i.e., chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e., sequence, structure, and pathway) information.</td>
<td>766,920 triples; 4,800 drugs; 2,500 protein sequences</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>LinkedCT</td>
<td>Clinical Trials</td>
<td>Linked data source of trials from ClinicalTrials.gov</td>
<td>7M triples; 62000 trials</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Medicare</td>
<td>Medicare Formulary</td>
<td>List of drugs that recipients of Medicare D are eligible to receive.</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Patient Records</td>
<td>Patient Data</td>
<td>Hand-generated test patient data, assuming data was collected within a PCHR (personally controlled health record).</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>PharmGKB</td>
<td>Genetic Information / Drug Response</td>
<td>Contains information that relates genetic variation to variation in drug response.</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>RDF-TCM</td>
<td>Genes / Diseases / Medicines / Ingredients</td>
<td>Traditional Chinese medicine, gene and disease association dataset and a linkset mapping TCM gene symbols to Extrez Gene IDs created by Neurocommons.</td>
<td>117,643 triples</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>SIDER</td>
<td>Diseases / Side Effects</td>
<td>SIDER contains information on marketed drugs and their adverse effects.</td>
<td>192,515 triples; 1,737 genes</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>STITCH</td>
<td>Chemicals / Proteins</td>
<td>STITCH contains information on chemicals, proteins, and their interactions.</td>
<td>7,500,000 chemicals; 500,000 proteins; 370 organisms</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
TMO
Sample Query

Which existing marketed drugs might potentially be re-purposed for AD because they are known to modulate genes that are implicated in the disease?

<table>
<thead>
<tr>
<th>drug_name</th>
<th>disease2_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s)-rolipram</td>
<td>Schizophrenia</td>
</tr>
<tr>
<td>(s)-rolipram</td>
<td>Autistic Disorder</td>
</tr>
<tr>
<td>(s)-rolipram</td>
<td>Bipolar Disorder</td>
</tr>
<tr>
<td>(s)-rolipram</td>
<td>Depression</td>
</tr>
<tr>
<td>irbesartan</td>
<td>Hypertension</td>
</tr>
<tr>
<td>lisinopril</td>
<td>Hypertension</td>
</tr>
<tr>
<td>lisinopril</td>
<td>Diabetes Mellitus, Insulin-Dependent</td>
</tr>
<tr>
<td>nifedipine</td>
<td>Hypertension</td>
</tr>
<tr>
<td>perindopril</td>
<td>Proteinuria</td>
</tr>
<tr>
<td>perindopril</td>
<td>Diabetes Mellitus, Non-Insulin-Dependent</td>
</tr>
<tr>
<td>perindopril</td>
<td>Cerebrovascular Accident</td>
</tr>
<tr>
<td>perindopril</td>
<td>Cardiovascular Diseases</td>
</tr>
<tr>
<td>perindopril</td>
<td>Dementia</td>
</tr>
<tr>
<td>perindopril</td>
<td>Hypertension</td>
</tr>
<tr>
<td>perindopril</td>
<td>Memory Disorders</td>
</tr>
<tr>
<td>pravastatin</td>
<td>Coronary Arteriosclerosis</td>
</tr>
</tbody>
</table>
TMO

• home:
  – http://esw.w3.org/topic/HCLSIG/PharmaOntology

• source code / TMO:
  – http://www.w3.org/2001/sw/hcls/ns/transmed
  – http://code.google.com/p/translationalmedicineontology/

• data sources (text search & SPARQL endpoint):
  – http://tm.semanticscience.org/fct
  – http://tm.semanticscience.org/sparql

• example queries:
  – http://esw.w3.org/topic/HCLSIG/PharmaOntology/Queries
Summary & Next Steps

- Strenghts
  - lots of free pharma/drug/translational medicine relevant data have been made available in a very flexible form
  - a first TMO candidate has been developed

- Weaknesses
  - non-techie interfaces to data and tailored applications building on linked data sets and ontologies are needed
  - lack of freely available clinical data

- Future TMO work:
  - tighter ontology/data integration
  - revisit mapping procedures
  - flexible integration of candidate domain ontologies/vocabularies
  - interfaces
Acknowledgements

• TMO
  – Colin Batchelor, Christine Denney, Christopher Domarew, Michel Dumontier, Anja Jentzsch, Joanne Luciano, Susie Stephens, Patricia L. Whetzel
  – Bosse Andersson, Olivier Bodenreider, Tim Clark, Lee Harland, Vipul Kashyap, Peter Kos, Julia Kozlovsky, James McGurk, Chimezie Ogbuji, Eric Prud'hommeaux, Matthias Samwald, Lynn Schriml, Jun Zhao

• LODD
  – Bosse Anderssen, TN Bhat, Chris Bizer, Don Doherty, Michel Dumontier, Anja Jentzsch, Oktie Hassanzadeh, Scott Marshall, Glen Newton, Eric Prud'hommeaux, Matthias Samwald, Susie Stephens, Kristin Tolle, Egon Willighagen, Jun Zhao
  – Eli Lilly

• W3C / Semantic Web for Health Care and Life Sciences Interest Group