The Translational Medicine Ontology
A Small Compass for Navigating a Large Sea of Biomedical Data

Elgar Pichler
W3C HCLSIG TMO Team Members

CSHALS 2010, Cambridge, MA
February 25, 2010
Outline

• Questions & Problems

• Translational Medicine Ontology (TMO)
  – Ontology
  – Data
  – Examples

• General Comments
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
Questions & Problems
Aspirin – nothing new, right?

New recommendations for cardiovascular disease prevention with Aspirin:

- slightly lower daily dose than baby aspirin
- yes for person with risk factors but no history of bleeding and ulcers; for men >45y, women >55y
- no for men <45y, women <55y, or >80y

• New findings every day.
• How does this affect the use of a drug? How does it affect me?
TMO Mission

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.
TMO Development

Concept Identification via Use Cases

Process:

– describe roles, work out use cases
– identify used concepts
– map concepts to other ontologies/vocabularies
– align with Basic Formal Ontology (BFO)
– identification of candidate domain ontologies
– refine and start over again

[bottom-up approach; compare tutorial by John Madden]
## Roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Primary Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellular and molecular biologists</td>
<td>Assessing target viability</td>
</tr>
<tr>
<td>Cheminformatician</td>
<td>Analyzing chemical data and making predictions</td>
</tr>
<tr>
<td>Clinical decision support</td>
<td>Analyzing response to therapies</td>
</tr>
<tr>
<td>Clinical trial formulator</td>
<td>Designing clinical trials</td>
</tr>
<tr>
<td>Health plan provider</td>
<td>Providing insurance coverage to individuals</td>
</tr>
<tr>
<td>Immunologist</td>
<td>Developing large molecules for therapeutic purposes</td>
</tr>
<tr>
<td>In vitro biologist</td>
<td>Predicting success of compounds to be tested in vivo</td>
</tr>
<tr>
<td>In vivo biologist</td>
<td>Performing toxicology and efficacy studies in animals</td>
</tr>
<tr>
<td>Medicinal chemist</td>
<td>Exploring structural patterns and properties of compounds</td>
</tr>
<tr>
<td>Primary care clinician</td>
<td>Treating broad range of patients</td>
</tr>
<tr>
<td>Project manager</td>
<td>Prioritizing activities and resources</td>
</tr>
<tr>
<td>Sales and marketing</td>
<td>Driving sales</td>
</tr>
<tr>
<td>Specialty medical provider</td>
<td>Treating patients with specific diseases</td>
</tr>
<tr>
<td>Statistician</td>
<td>Testing scientific hypotheses using statistical approaches</td>
</tr>
<tr>
<td>Strategic/portfolio manager</td>
<td>Assessing market opportunities</td>
</tr>
<tr>
<td>Systems physiologist</td>
<td>Understanding the biological system</td>
</tr>
</tbody>
</table>
TMO Development
Concept Identification via Use Cases

Example

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


[...]

Elgar Pichler  
CSHALS 2010  
Translational Medicine Ontology  
8
TMO Development
Mapping to Other Ontologies/Vocabularies

NCBO
TMO Development
Mapping to Other Ontologies/Vocabularies

UMLS
## TMO Development
### 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

Use of Other Ontologies/Vocabularies

Ontologies used in TMO:

- Experimental Factor Ontology (EFO): cell line
- Information Artifact Ontology (IAO): class annotations
- Ontology for Biomedical Investigations (OBI): planned process, label molecular entity, metabolite
- Protein Ontology (PRO): protein
- Sequence Ontology (SO): SNP, gene, copy number variation, genotype
TMO Development
Alignment with BFO

<100 main TMO classes
aligned with BFO
TMO

Data Aggregation

Process:

– rdf-ize data
– load data into Virtuoso triple store
– generate mappings (sameAs links) between data sources and TMO via
  – same IDs
  – string & semantic matching (LinQuer, SILK)
## TMO

### Data Sources

<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>x</td>
<td></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></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></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></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></td>
<td></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></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></td>
<td>x</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 Data Mapping to TMO

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Mapping (Data Source to TMO)</th>
</tr>
</thead>
</table>
| ClinicalTrials.gov       | - '??' maps to 'textual entity' (IAO_0000300)  
- 'diagnostic criteria' maps to 'diagnostic criterion' (TMO_0068)  
- 'diagnostic criteria' maps to 'diagnostic inclusion criterion' (TMO_0069)  
- 'diagnostic criteria' maps to and 'diagnostic exclusion criterion' (TMO_0070) |
| Diagnostic Data (DuBois) | - 'drugs' map to 'pharmaceutical product' (TMO_0002)  
- 'ingredients' map to 'active ingredient'  
- 'organization' maps to 'institution' (TMO_0025) |
| DailyMed                 | - 'disease' maps to 'disease' (TMO_0047)  
- 'gene' maps to 'gene' (SO:0000704) |
| Diseasome                | - 'drug-drug interactions' maps to 'drug-drug interaction' (TMO_0040)  
- 'drugs' maps to 'pharmaceutical product' (TMO_0002)  
- 'targets' map to 'target' (TMO_0006) |
| DrugBank                 | - 'drugs' map to 'pharmaceutical product' (TMO_0002) |
| Patient Records          | - 'association' maps to 'study result' (OBI_0000682) |
| PharmGKB                 | - 'drugs' map to 'active pharmaceutical ingredient' (TMO_0000)  
- 'side effects' map to 'adverse drug event' (TMO_0043) |
| SIDER                    | - '??' maps to 'textual entity' (IAO_0000300)  
- 'diagnostic criteria' maps to 'diagnostic criterion' (TMO_0068)  
- 'diagnostic criteria' maps to 'diagnostic inclusion criterion' (TMO_0069)  
- 'diagnostic criteria' maps to and 'diagnostic exclusion criterion' (TMO_0070) |
Sample Queries ... and Answers

Discovery:

_ What genes are associated with or implicated in AD?
At least 97 genes have some association with AD.

_ Which existing marketed drugs might potentially be re-purposed for AD because they are known to modulate genes that are implicated in the disease?
57 compounds or classes of compounds that are used to treat 45 diseases.

Physician

_ What are the diagnostic criteria for AD?
12 Diagnostic inclusion criteria and 9 exclusion criteria were obtained from the criteria outlined in Dubois et al.

_ Is Donepezil covered by Medicare D?
Yes, Medicare D covers two brand name formulations of Donepezil.

Clinical:

_ What active trials are ongoing that would be a good fit for Patient 2?
58 Alzheimer trials, 2 mild cognitive impairment trials, 1 hypercholesterolaemia trial, 66 myocardial infarction trials, 46 anxiety trials, and 126 depression trials.
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
Comments

Personal Wishlist – Manifesto-Worthy?

- Data providers should make their data available in RDF, with SPARQL endpoints.
- SuperMapper should do all of our mapping work, specify which kind of mapping is used, and record relevant provenance data.
- Federated query should be enabled with access policy mediation.

[See also comments on sameAs and provenance in talk by James McCusker & Deborah Mc Guinness.]
[Talk to Eric Prud'hommeaux about concept for access policy mediation for SPARQL endpoints.]
Summary

Plus

- Several pharma/drug/translational medicine relevant data are available as linked data set.
- A first TMO candidate has been developed.
- The TMO project is a great example of a collaboration between industry, academia, and W3C HCLS in the pre-competitive space.

Minus

- More intuitive and tailored interfaces to linked data are needed.
- There is a lack of freely available clinical data.

Future TMO work:

- Tighter ontology/data integration.
- Revisit mapping procedures.
- Flexible integration of candidate domain ontologies/vocabularies.
- Interfaces.

[Data & know-how sharing, compare also talk by Vijay Bulusu]
[W3C HCLSIG, see talks from Tutorial session and by Susie Stephens]
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.

http://esw.w3.org/topic/HCLSIG/LODD

LODD data in the Linked Data cloud ... 

... are represent 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

• 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/loodd/topquestions/
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